UNIVERSITÉ DE MONTPELLIER
École doctorale Information Structures Systèmes (I∾S)

MÉMOIRE DE SYNTHÈSE

en vue d'une candidature à une

HABILITATION À DIRIGER DES RECHERCHES

par

Frédéric Mortier

Modélisation statistique multivariée pour l'écologie et la génétique.

Soutenue le 09 juin devant la commission d'examen :

M.	Ali Arab	Rapporteur
M.	Jean-Noël Bacro	Président
M.	Olivier Gimenez	Examinateur
M.	Frédéric Gosselin	Rapporteur
Mme.	Marie-Laure Martin-Magniette	Rapporteur
M.	Étienne Rivot	Examinateur

C'est en essayant continuellement que l'on finit par réussir. En d'autres termes : plus ça rate, plus on a de chances que ça marche
(Proverbe Shadokien)

C'est en forgeant que l'on devient musicien
(Proverbe Shadokien)

Résumé Le devenir des forêts est désormais l'une des préoccupations majeures du $20^{\text{ième}}$ siècle. Celles-ci sont justifiées par l'importance que revêtent les forêts pour de multiples acteurs et à de multiples échelles. L'enjeu consiste aujourd'hui à conserver la biodiversité des forêts tropicales et à les gérer durablement, c'est-à-dire à exploiter leurs ressources en préservant à long terme leurs fonctions écologiques, économiques et sociales. Protéger et gérer durablement un écosystème dans son ensemble conduit à le considérer non plus comme un ensemble indépendant de processus biologiques mais comme un ensemble de processus interdépendants. Analyser, comprendre ou encore prédire le future de ces écosystèmes nécessite certaines précautions et des méthodes d'analyses adéquates doivent être employées. C'est ce que je me suis efforcé de faire au cours de ma carrière et ce mémoire, d'habilitation à diriger des recherches, présente les travaux que j'ai été amenés à développer. Il est important de souligner que ce sont les questions biologiques qui ont motivé mes recherches en statistique. Il m'est donc apparu naturel que ce soit au travers des applications que je devais présenter mes activités de recherches en bio-statistiques. La première partie donne un rapide aperçu du contexte biologique et mathématique. La seconde présente plus en détail quatre résultats qui me semblent majeurs et qui traitent de la prise en compte des dépendances spatiales, de la richesse spécifique des écosystèmes tropicaux ou encore des questions de prédictions. La dernière partie présente les stratégies à long terme que je souhaiterais mettre en place pour mener à bien et fédérer les recherches et répondre ainsi à l'objectif commun : la préservation des écosystèmes forestiers compatible avec le développement des populations humaines.

Table des matières

Première partie

Présentation du candidat

1 Curriculum vitæ

Frédéric Mortier
Né le 16 décembre 1971
Nationalité française
Marié

29 bis, rue Lakanal
34000 Montpellier
Tél. : 04 67 03 30 05

Cirad, département « Environnements et Sociétés »
Campus de Baillarguet, TA C-105/D
34 398 Montpellier Cedex 5
Tél. : +33 (0)4 67 59 37 66
Fax : +33 (0)4 67 59 37 33
E-mail : fmortier@cirad.fr

1.1 Titres

oct. 1999–déc. 2002 **Thèse de doctorat** de l'Université Lille I, intitulée « Estimation de distances généralisées – Application à la distinction variétale »
Directeur de thèse : Avner Bar-Hen (Univ. Paris Descartes).
Co-directeur de thèse : Stéphane Robin (UMR518 AgroParisTech/INRA) Laboratoire d'accueil : Laboratoire Paul Painlevé, Mathématiques et Informatique Appliquée.
Co-directeur de thèse : Claire Baril (GEVES)

1998–1999 DEA BioStatistique de l'Université Montpellier II

1997–1998 Maîtrise Mathématiques, Université Denis Diderot, Paris

1995–1997 DEUG A et Licence Mathématiques, Université Denis Diderot, Paris VII

1.2 Fonction occupée

2003–2010 **Chercheur** au Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), département forêts, UPR 37, « Génétique et Amélioration des espèces forestières »

2010– **Chercheur** au Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), département Environnements et Sociétés, UPR 105, « Biens et Services des Écosystèmes Forestiers (B&sef) »

1.3 Formations et enseignements dispensés

1. Linear Mixed Models, theory and applications. Cours et TD (20h) *Master BioStatistques*, Montpellier, France. Avec C. Trottier. Dispensé en 2012–2013 et 2013–2014.
2. Initiation au logiciel **R**. Cours et TD (20h) *CIRAD*, Montpellier, France. Avec G. Cornu et H. Dessard. Dispensé en 2013.
3. Initiation au logiciel **R**. Cours et TD (20h) *CIRAD*, Montpellier, France. Avec G. Cornu et H. Dessard. Dipensé en 2012.
4. Linear Mixed Models, theory and applications. Cours (6 h). *ESALQ*, Piracicaba, Brésil. Avec C. Trottier. Dispensé en 2012.
5. Initiation au logiciel **R**. Cours et TD (24h). *Embrapa*, Bélem, Brésil. Avec F. Wagner. Dispensé en 2012.
6. Initiation au logiciel **R**. Cours (18h). *CIRAD*, Montpellier, France. Avec M. Denis et F. Ribierre. Dispensé en 2011.
7. Generalized Linear Mixed Models, theory and applications. Cours et TD (36 h). *CIRAD*, Montpellier, France. Avec C. Trottier et C. Demetrio. Dispensé en 2011.

8. Generalized Linear Mixed Models, theory and applications. Cours et TD (36 h). *ESALQ*, Piracicaba, Brésil. Avec C. Trottier et C. Demetrio. Dispensé en 2009.

1.4 Participation

1.4.1 à des projets

1. Fonds Français pour l'Environnement Mondial (FFEM) et l'Agence Française de Développement (AFD). *Dynamique des forêts d'Afrique centrale (DynAfFor).* Coord S. Gourlet-Fleury (2013–2018).
2. ERA-Net BiodivERsA. *Congo Forest Tipping Points (CoForTips).* Coord : C. Garcia. (2013–2015)
3. FEDER-ANR. *Guyasim.* Coord : V. Rossi (2011–2013).
4. ERA-Net BiodivERsA. *Congo Forest Change (CoForChange).* Coord : S. Gourlet-Fleury (2009–2012).
5. INCO . *INNOVKAR.* Coord : J.M. Bouvet (2006–2009).
6. Projet national : BRG. *Un modèle de variabilité fonctionnelle chez les arbres forestiers : le gène CCR d'eucalyptus.* Coord : J.M Gion (2005-2007).

1.4.2 à des événements

1. Membre du conseil d'organisation de « International Statistical Ecology Conference 2014 ».
2. Organisation d'une session « Modèles hiérarchiques spatiaux-temporels pour l'étude de la dynamique des populations » au congrès *Écologie 2010*. Lancé à l'initiative des réseaux ComEvol, EcoVeg, GDR Traits, JEF, PPD, et REID. Avec Nicolas Bez.
3. Organisation des journées « Amélioration des plantes », *CIRAD* 2008.

1.5 Divers

– Relecteur pour les revues suivantes : *Annals of Forest Science*, *Bois et Forêts des Tropiques*, *Entropy*, *Ecological Modelling*, *Forest Ecology and Management*, *Journal of Agricultural, Biological, and Environmental Statistics*, *Journal of The Royal Statistitical Society, series C*, *Molecular Ecology*, *Tree Genetics and Genomics*.

- Membre du groupe « Statistiques et Environnement » de la Société Française de Statistiques.
- Membre du Groupe de Recherche « Écologie Statistique », lancé à l'initiative d'Olivier Gimenez.
  - Responsable du thème de recherche *Ressources renouvelables*. Avec V. Trenkel.
  - Responsable de la question transversale *Selection de modèle*. Avec Olivier Gimenez.
- Membre du jury de recrutement d'un biologiste spécialisé en écologie forestière pour un poste de chercheur au « département Environnements et Sociétés du Cirad ».
- Membre du jury de recrutement d'un maître de conférence à l'université Montpellier III en statistiques et probabilités.

2
Liste des publications

Les listes sont données par ordre chronologique.

2.1 Méthodes

1. *F. Mortier*, D.-Y. Ouédraogo, F. Claeys, M.G. Tadesse, G. Cornu, F. Baya, F. Benedet, V. Freycon, S. Gourlet-Fleury and N. Picard. Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics*. 26 : 39– 51, 2015.
2. X. Bry, C. Trottier, T. Verron and *F. Mortier*. Supervised Component Generalized Linear Regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119(0) : 47–60, 2013.
3. V. Garreta, J. Guiot, *F. Mortier*, J. Chadoeuf and C. Hély. Pollen-based climate reconstruction : Calibration of the vegetation-pollen processes. *Ecological Modelling*, 235-236 : 81–94, 2013.
4. *F. Mortier*, V. Rossi, G.. Guillot, S. Gourlet-Fleury and N. Picard. Population dynamics of species-rich ecosystems : the mixture of matrix population models approach. *Methods in Ecology and Evolution*, 4(4) : 316–326, 2012.
5. P. Chagneau, *F. Mortier*, N. Picard and J.N. Bacro. Prediction of a multivariate spatial random field with continuous, count and ordinal outcomes. *Biometrics*, 58(3) : 345–367, 2011.

6. N. Picard, *F. Mortier*, L. Saint-André, C. Trotta and M. Henry. Using Bayesian model averaging to predict tree aboveground biomass. *Forest Science*, 58(1) : 15–23, 2011.
7. P. Chagneau, *F. Mortier* and N. Picard. Designing permanent sample plots by using a spatially hierarchical matrix population model. *Journal of the Royal Statistical Society Series C*, 58(3) : 345–367, 2009.
8. N. Picard, *F. Mortier* and P. Chagneau. The multi-scale marked area interaction point processes : a model for the spatial pattern of trees. *Scandinavian Journal of Statistics*, 36(1) : 23–41, 2009.
9. O. Flores, V. Rossi and *F. Mortier*. Autocorrelation offsets zero-inflflation in models of tropical saplings density. *Ecological Modelling*, 220(15) : 1797–1809, 2009.
10. F. Chaubert, *F. Mortier* and L. Saint André. Multivariate dynamic model for ordinal outcomes. *Journal of Multivariate Analysis*, 99(8) : 1717–1732, 2008.
11. N. Picard, P. Chagneau, *F. Mortier* and A. Bar-Hen. Predicting the stock recovery rate of a tropical species using matrix models. *Ecological Modelling*, 214 : 349–360, 2008.
12. *F. Mortier*, S. Robin, S. Lassalvy, C.P. Baril and A. Bar-Hen. Prediction of Euclidean distances with discrete and continuous outcomes. *Journal of Multivariate Analysis*, 97(8) : 1799 – 1814, 2006.
13. G. Guillot, A. Estoup, *F. Mortier* and J.F. Cosson. A spatial statistical model for landscape genetics. *Genetics*, 170 : 1261–1280, 2005.
14. A Bar-Hen and *F. Mortier*. Influence and sensitivity measures in correspondence analysis. *Statistics*, 38(3) : 207–215, 2004.

2.2 Applications

1. F. Kleinschroth, S. Gourlet-Fleury, P. Sist, *F. Mortier* and J.R.Healey. Legacy of logging roads in the Congo Basin : how persistent are the scars in forest cover ? *Ecosphere*, 6(4) :art64, 2015.
2. O. Gimenez, S.T. Buckland, B.J.T. Morgan, N. Bez, S. Bertrand, R. Choquet, S. Dray, M.-P. Etienne, R. Fewster, F. Gosselin, B. Mérigot, P. Monestiez, J. M. Morales, *F. Mortier*, F. Munoz, O. Ovaskainen, S. Pavoine, R. Pradel, F.M. Schurr, L. Thomas, W. Thuiller, V. Trenkel, P. de Valpine, E. Rexstad. Statistical ecology comes of age. *Biology Letters*, 10 (12) : 20140698, 2014.

3. E. Mandrou, M. Denis,C. Plomion, F. Salin, F. Mortier and J.-M. Gion,. Nucleotide diversity in lignification genes and QTNs for lignin quality in a multi-parental population of Eucalyptus urophylla. *Tree Genome and Genetics*, 10(5) : 1281–1290, 2014.
4. D.-Y. Ouédraogo, *F. Mortier*, S. Gourlet-Fleury, V. Freycon and N. Picard. Slow-growing species cope best with drought : evidence from long-term measurements in a tropical semi-deciduous moist forest of Central Africa. *Journal of Ecology*. 101(6) : 1459–1470, 2013.
5. S. Gourlet-Fleury, D. Beina, A. Fayolle, D.-Y. Ouédraogo, *F. Mortier*, F. Bénédet, D. Closset-Kopp and G. Decocq. Silvicultural disturbance has little impact on tree species diversity in a Central African moist forest. *Forest Ecology and Management*. 304 : 322–332, 2013.
6. S. Gourlet-Fleury, *F. Mortier*, A. Fayolle, D.-Y. Ouédraogo, F. Bénédet and N. Picard, Will moist forests recover from logging in Central Africa ? Insights from a longitudinal long-term silvicultural experiment. *Philosophical Transactions B*. 368 : 20120302, 2013.
7. N. Picard, P. Köhler, *F. Mortier* and Gourlet-Fleury S. A comparison of five classifications of species into functional groups in tropical forests of French Guiana. *Ecological complexity*, 11 : 75–83, 2012.
8. A. Fayolle, B. Engelbrecht, V. Freycon, *F. Mortier*, M. Swaine, M. Réjou-Méchain, J-L. Doucet, N. Fauvet, G. Cornu, and S. Gourlet-Fleury. Geological substrates shape tree species and trait distributions in African moist forests. *Plos One*, 7(8) : e42381-e42381, 2012.
9. H. Wernsörfer, H. Caron, S. Gerber, G. Cornu, V. Rossi, *F. Mortier* and S. Gourlet-Fleury. Relationships between demography and gene flow and their importance for the conservation of tree populations in tropical forests under selective felling regimes. *Conservation Genetics*, 12(1) : 15–29, 2011.
10. E. Mandrou, *F. Mortier*, M. Denis, G. Chaix, E. Villar, C. Plomion and J.-M. Gion. Functional variability of two lignification genes in eucalyptus urophylla. *BMC Proceedings*, 5(Suppl 7) : O12, 2011.
11. D.Y. Ouédraogo, D. Beina, N. Picard, *F. Mortier*, F. Baya and S. Gourlet-Fleury. Thinning after selective logging facilitates floristic composition recovery in a tropical rain forest of Central Africa. *Forest Ecology and Management*. 262(12) : 2176–2186, 2011.
12. N. Picard, A. Bar-Hen, *F. Mortier* and J. Chadoeuf. Understanding the dynamics of an undisturbed tropical rain forest from the spatial pattern of trees. *Journal of Ecology*, 97(1), 97–108, 2009.

13. G. Guillot, *F. Mortier* and A. Estoup. Geneland : A program for landscape genetics. *Molecular Ecology Ressources*, 5 : 712–715, 2005.

2.3 Actes de congrès

1. X. Bry, C. Trottier, T. Verron, *F. Mortier*, Supervised Component Generalized Linear Regression using a PLS-extension of the Fisher scoring algorithm, *20^{th} International Conference on Computational Statistics (COMPSTAT)* 121-129, 2012.
2. K. Trévennec, *F. Mortier*, F. Lyazrhi, H. Thu Huong, V. Chevalier, and F. Roger. Swine influenza in vietnam : preliminary results of epidemiological studies. *influenzaand other respiratory viruses*. Hong Kong, Chine. 5 : 71–73, 2011.
3. P. Chagneau, *F. Mortier*, Nicolas Picard, and J-N Bacro. Hierarchical bayesian model for gaussian, poisson and ordinal random fields. *geoEBV VII - Geostatistics for Environmental Applications*, volume 16 of Quantitative Geology and Geostatistics, Southampton, UK. pages 333–344, 2010.
4. P. Chagneau, *F. Mortier*, Nicolas Picard, and Jean-Noël Bacro. Prediction of a multivariate spatial random field with continuous, count and ordinal outcomes. *Proceedings of the Eight International Geostatistic Congress, volume 16 of GEOSTAT*, Santiago, Chili. 1 : 479–488, 2008.
5. *F. Mortier*, O. Flores and S. Gourlet-Fleury. Spatial bayesian models of tree density with zero inflation and autocorrelation. *Journal de la Société française de statistique et Revue de statistique appliquée*, Paris, France. 148(1) : 39–51, 2007.

2.4 Conférences

1. F. Claeys, *F. Mortier*, D.-Y. Ouédrogo, L. François, B. Herault, R. Gaspard, A. Fayolle, N. Picard, M.G. Tadesse, S. Gourlet-Fleury. Predicting the combined impacts of climate change and selective logging in production forests of Central Africa. *Our Common Future under Climate Change (OCFCC) UNESCO*, Paris, France, 2015.
2. *F. Mortier*, D.-Y. Ouédraogo, F. Claeys, M.G. Tadesse, G. Cornu, F. Baya, F. Benedet, V. Freycon, S. Gourlet-Fleury, and N. Picard. Mixture of inhomogeneous matrix models for species-rich ecosystems. *George Washington University*, Washington D.C., USA, 2015. (Invité).

3. *F. Mortier*, D.-Y. Ouédraogo, F. Claeys, M.G. Tadesse, G. Cornu, F. Baya, F. Benedet, V. Freycon, S. Gourlet-Fleury, and N. Picard. Mixture of inhomogeneous matrix models for species-rich ecosystems. *American Mathematical Society*, session on *Characterizing Uncertainty for Modeling Physical Processes*, Washington D.C., USA , 2015. (Invité).
4. *F. Mortier*, D.-Y. Ouédraogo, F. Claeys, M.G. Tadesse, G. Cornu, F. Baya, F. Benedet, V. Freycon, S. Gourlet-Fleury, and N. Picard. Mixture of inhomogeneous matrix models for species-rich ecosystems. *Eastern North American Region International Biometric Society*, session on *Recent Advances in Statistical Ecology*, Miami, USA, 2015. (Invité).
5. *F. Mortier*, D.-Y. Ouédraogo, F. Claeys, M.G. Tadesse, G. Cornu, F. Baya, F. Benedet, V. Freycon, S. Gourlet-Fleury, and N. Picard. Mixture of inhomogeneous matrix models for species-rich ecosystems. *Rice University*, Huston, TX, USA, 2014. (Invité).
6. *F. Mortier*, X. Bry, G. Cornu, C. Trottier. SCGLR : A component-based multivariate regression method to model species distributions. *International Statistical Ecology Conference (ISEC)*. Montpellier, France, 2014.
7. *F. Mortier*. Impact of anthropogenic and climatic changes on biomass and diversity of Central African forests, from local to global scale : original methods for new results. *European Geoscience Union*. Vienne, Autriche, 2014. (Invité).
8. C. Trottier, X. Bry, G. Cornu, *F. Mortier*. SCGLR : Un package R pour la régression linéaire généralisée sur composantes supervisées. *3ème rencontres R*. Montpellier, France, 2014.
9. X. Bry, C. Trottier, T. Verron, *F. Mortier*. Supervised Component Generalized Linear Regression using a PLS-extension of the Fisher scoring algorithm. *45e Journées de Statititstiques de la SFDS* , Toulouse, France, 2013.
10. *F. Mortier*, V. Rossi, G. Guillot, S. Gourlet-Fleury and N. Picard. Population dynamics of species-rich ecosystems : the mixture of matrix population models approach.*Statistiques Appliquées au Developpement en Afrique (SADA)* , Cotonou, Benin, 2013.
11. *F. Mortier*, D. Ouedraogo, and N. Picard. Finite mixture of size projection matrix models for highly diverse rainforests in a variable environment. *In 57 Reuniao Anual da RBras*. Piracicaba, Brésil, 2012. (Invité)
12. *F. Mortier*, P. Chagneau, M.P. Etienne, N. Picard, C. Piou, and Rossi V. Modélisation bayésienne hiérarchique pour l'écologie et la recherche

environnementale. *In Journées de statistique de la Société Française de Statistique.* Grammath, Tunisie, 2011. (Invité)

13. D. Ouédraogo, *F. Mortier*, and N Picard. Incorporating environmental variability in matrix models predictions for highly diverse rainforests. *In Research priorities in tropical silviculture : towards new paradigms ?* IUFRO International Conference., pages 15(18), Montpellier, France, 2011.
14. D. Ouédraogo, D. Beina, N. Picard, *F. Mortier*, F. Baya, and S. Gourlet-Fleury. Thinning after selective logging facilitates floristic composition recovery in a tropical rain forests of central africa. *In Research priorities in tropical silviculture : towards new paradigms ?* IUFRO International Conference, pages 15–18, Montpellier, France, 2011.
15. G. Vieilledent, S. Gourlet-Fleury, and *F. Mortier*. twoe : An R package for modelling tropical forest dynamics from permanent sample plots using a hierarchical bayesian approach to capture species diversity. *In Research priorities in tropical silviculture : towards new paradigms ?* IUFRO International Conference, Montpellier, France, 2011.
16. P. Chagneau, *F. Mortier*, N. Picard and J-N. Bacro. Processus de Cox marqé dirigé par un environnement prédit : application à la répartition spatiale de juvéniles en forêt tropicale humide. *41èmes Journées de la Société Française de Statistiques*, Bordeaux, France, 2009.
17. V. Garreta, *F. Mortier* and J. Chadoeuf. Modéliser le pollen piégé au sol en fonction de la végétation simulée par LPJ-GUESS : Un modèle hiérarchique des processus intégrant sur-dispersion et zéros structurels. *41èmes Journées de la Société Française de Statistiques*, Bordeaux, France, 2009.
18. *F. Mortier*, V. Rossi, N. Picard and S. Gourlet-Fleury. Unsupervised classiffication of species groups based on mixture matrix population models. *In Modélisation des Ecosystèmes Tropicaux et Amazoniens (META)*, Kourou, France, 2007.
19. P. Chagneau, *F. Mortier* and N. Picard. Asymptotic properties of spatially hierarchical matrix population models. *39èmes Journées de la Société Française de Statistiques*, Angers, France, 2007.
20. F. Chaubert and *F. Mortier*. Modèle Probit Multivarié Ordinal Dynamique. Application à l'estimation de la Biomasse d'un peuplement forestier d'eucalyptus. *38èmes Journées de la Société Française de Statistiques*, Clamart, France, 2006.
21. A. Bar-Hen and *F. Mortier* : Mesure d'influence en analyse factorielle

des correspondances. *32ème journées de statistique de la Journées de la Société Française de Statistiques*, Fès, Maroc, 2000.

2.5 `R` Packages

1. Co-construction
   - SCGLR : Supervised Component Generalized Linear Regression. (With G. Cornu, C. Trottier et X. Bry).
   - Genland : A Spatial Statistical Model for Landscape Genetics, version 1. (With G. Guillot, A. Estoup).
2. Contribution
   - FLXMRglmnet : FlexMix Interface for Adaptive Lasso / Elastic Net with GLMs (With N. Picard).

3
Encadrement d'étudiants

– 6 thèses.
– 7 stages de M2 recherche.

3.1 Thèses

3.1.1 Co-directeur de thèse

Alexandra Jestin (2015–) effectue sa thèse au sein de l'équipe « Biens et Services des Écosystèmes Forestiers » (B&SEF), de l'UMR « Amélioration génétique et adaptation des plantes méditerranéennes et tropicales » (AGAP) sur la question *Sélection de variables pour données longitudinales en mélange avec effets différentiels dans le temps : application à la modélisation multi-spécifique et à l'amélioration génétique.* Cette thèse est co-encadrée par J-N. Bacro de l'université de Montpellier et M. Denis de l'UMR AGAP.

Romain Gaspard (2014–2017) effectue sa thèse au sein de l'équipe « Biens et Services des Écosystèmes Forestiers » et l'UMR « ECologie de FOrêts de Guyanne » sur la question : *l'exploitation forestière rend-elle les forêts tropicales plus vulnérables aux changements climatiques ?* Cette thèse est co-encadrée par B. Herault de l'UMR ECoFoG et S. Gourlet-Fleury de l'UR B&SEF.

Pierrette Chagneau (2006-2009) a effectué sa thèse au sein de l'équipe « Dynamique des écosystèmes forestiers tropicaux » sur la question de la *Prédiction de la répartition spatiale de différents stades d'arbres en forêts tropicale humides à l'aide de processus ponctuels hétérogénes.* Co-encadrée par J.N Bacro de l'université Montpellier II et N. Picard du CIRAD, la thèse a été soutenue le 4 décembre 2009 à l'école doctorale « Information, Structures et Systèmes (I∾S) ». Pierrette Chagneau est actuellement maître de conférences en mathématiques appliquées à l'INSA de Rennes. Cette collaboration a donné lieu à trois articles.

Publications : P. Chagneau, et al. (2011) ; N. Picard, et al.. (2009) et N. Picard, et al. (2008)

Ciré Elimane Sall (2005-2009) a effectué sa thèse au sein de l'équipe « Amélioration génétique des espèces forestières ». Ciré Sall a travaillé sur la question de *l'estimation de l'apparentement par la méthode du maximum de vraisemblance composite avec prise en compte de la dépendance spatiale entre les individus*. Ce travail, co-encadré par A. Ganoun de l'université Montpellier II, a été soutenu le 22 octobre 2009 à l'école doctorale « Information, Structures et Systèmes (I∾S) ». Ciré E. Sall est chercheur à l'Institut Sénégalais de Recherche Agricole (ISRA). Sa thèse s'est déroulée en alternance entre les deux pays avec une présence en France de 4 à 5 mois par an.

3.1.2 Co-encadrement de thèse

Florian Claeys (2013–2016) effectue sa thèse au sein de l'équipe « Biens et Services des Écosystèmes Forestiers » sur la question de *l'amélioration de la durabilité et de la rentabilité de l'exploitation forestière tropicale par des instruments incitatifs*. Il est inscrit à l'école doctorale « Agriculture Alimentation BIologie Environnement Santé (ABIES) ». Cette thèse est co-encadrée par A. Karsenty et S. Gourlet Fleury du CIRAD et par P. Delacotte du laboratoire d'Économie Forestière (Lef) de l'Inra et AgroParisTech, Nancy. Ce travail a d'ores et déjà donné lieu à un article.

Publications : F. Mortier, D.-Y. Ouédraogo, F. Claeys, et al.. (2015)

Dakis Ouédraogo (2009-2011) a effectué sa thèse au sein de l'équipe « Biens et Services des Écosystèmes Forestiers » sur la question de *la prédiction de la dynamique forestière à l'aide d'un modèle matriciel qui incorpore la variabilité de la réponse des espèces à l'environnement : Application dans une forêt tropicale humide semi-décidue d'Afrique centrale*. Cette thèse co-encadrée avec J.D. Lebreton du « Centre d'Écologie Fonctionnelle et Évolu-

tive (CEFE) » et N. Picard du CIRAD a été soutenue le 12 décembre 2011 à l'université Montpellier II à l'école doctorale « Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement (SIBAGHE) ». Dakis est actuellement post-doctorante à l'université de Gembloux en Belgique. Cette collaboration a donné lieu à trois articles :

Publications : F. Mortier, et al. (2015) ; D.-Y. Ouédraog et al. (2013) et D.Y. Ouédraogo, et al. (2011)

Olivier Flores (2005, 12 mois) a effectué sa thèse au sein de l'équipe « dynamique des écosystèmes forestiers tropicaux » entre 2001 et 2005 sur le thème du *Déterminisme de la régénération chez quinze espéces d'arbres tropicaux en Guyane française : les effets de l'environnement et de la limitation par la dispersion.* Cette thèse était encadrée par E. Garnier (CEFE) et S. Gourlet-Fleury (CIRAD). Pendant sa dernière année, j'ai encadré Olivier Flores afin d'élaborer un modèle spatial pour des données de comptage avec une sur-représentation de zéros. Olivier Flores est maintenant maître de conférence à l'université de la Réunion. Cette collaboration a donné lieu a un article :

Publication : Flores O., et al. (2009).

3.2 Master Recherche

F. Xiao (2015-2016). *Variable selection in mixture of multivariate generalized linear mixed effects models. Model species distributions in highly biodiverse ecosystems accounting for species interactions and spatial dependence.* Ce stage se déroule dans le cadre des travaux de fin d'étude de M. Xiao à Georgetown University, Washington D.C., USA. Ce travail est co-encadré par M. Tadesse.

H. Li (2015, 5 mois). *Models for Tropical Moist Forests : sampling strategies for high-dimensional mixture regression models.* Ce stage a eu lieu dans le cadre des travaux de fin d'étude de Mlle Li à Georgetown University, Washington D.C., USA. Ce travail a été co-encadré par M. Tadesse.

F. Claeys (2012, 4 mois). *Optimisation de scénarios d'exploitation forestiére et de séquestration de carbone pour une gestion durable des forêts d'Afrique centrale.* Master d'Économie du Développement Durable, de l'Environnement et de l'Énergie d'AgroParisTech. Avec A. Karsenty, S. Gourlet-Fleury.

F. Rollot (2009, 5 mois). *Modélisation de la croissance d'essences forestiéres tropicales prenant en compte la variabilité inter-spécifique.* Master biostatistiques de l'université Montpellier II. Avec V. Rossi.

N. Zougab (2008, 5 mois). *Développement d'un modèle de génétique quantitative dynamique. Prise en compte des compétitions inter-individuelles.* Master biostatistiques de l'université Montpellier II. Avec V. Rossi.

C. Centurion (2006, 5 mois). *Impact d'un schéma d'amélioration sur l'évolution des intéractions Génotype × Environnement. Application à la sélection de clones d'eucalyptus en Uruguay chez l'Eucalyptus.* Master Ressources Phytogénétiques et Interactions biologiques de l'université Montpellier II. Avec P. Vigneron.

Pierrette Chagneau (2006, 4 mois). *Optimisation sous contraintes spatiales ; application à la mise en place de parcelles permanentes de suivi des forêts tropicales humides.* Master recherche en biostatistiques de l'université Montpellier II. Avec N. Picard.

Publications : P. Chagneau, et al.(2009)

F. Chaubert (2004, 5 mois). *Modéle Probit Multivarié Ordinal Dynamique. Application à l'estimation de la Biomasse d'un peuplement forestier d'eucalyptus.* Master recherche en biostatistiques de l'université Montpellier II. Avec L. Saint-André.

Publications : F. Chaubert, et al. 2008

Alicia Rebollo (2003, 5 mois). *Développement d'un modéle de génétique quantitative. Prise en compte des corrélations spatiales et temporelles. Application à l'amélioration génétique des eucalyptus.* Master recherche en biostatistiques de l'université Montpellier II.

3.3 Licence ou équivalence

3.4 Comités de thèse

Florence Carpentier (2010). *Mesure de la dispersion du pollen et des graines à partir de marqueurs génétiques.* Directeur de thèse : Etienne Klein et Joël Chadoeuf, INRA Avignon.

Éric Mandrou (2010) *Variabilité fonctionnelle de gènes candidats de la lignification chez l'eucalyptus*. Directeur de thèse Christophe Plomion, INRA.

Emily Walker (2010) *Analyse de la répartition spatiale de l'effort de pêche à micro et méso échelles : Cas de la pêcherie thonière dans l'Océan indien*. Directeur de thèse : Nicolas Bez, IRD.

Florence Chaubert (2008) *Modèle multiphasiques*. Directeur de thèse : Yann Guedon, CIRAD et Chistian Lavergne, université Montpellier II.

Loraine Bottin (2006) *Déterminants de la variation moléculaire et phénotypique d'une espèce forestière en milieu insulaire Cas de Santalum austrocaledonicum en Nouvelle-Calédonie*. Directeur de thèse : Jean-Chistophe Glaszmann, CIRAD.

Deuxième partie

Synthèse des travaux de recherche et perspectives

J'ai été recruté au CIRAD en 2003 pour apporter, à mes collègues biologistes, généticiens et écologues, mon expertise en tant que statisticien appliqué. J'ai travaillé au sein de deux équipes, l'équipe « Diversité génétique et amélioration des espèces forestières » de 2003 à 2010 et depuis 2010 j'exerce mes activités au sein de l'équipe « Biens et Services des Écosystèmes Forestiers » (B&sef). Bien que les objectifs et les questions biologiques soient différents pour chacune des deux équipes, les outils méthodologiques nécessaires sont similaires ou très proches. Dans le cadre du programme d'amélioration génétique des espèces pérennes (eucalyptus, teck...), mesurer l'héritabilité [1] des caractères d'intérêt agronomique tels que la croissance en hauteur, l'aptitude au bouturage ou le nombre de boutures par pied mère par exemple, est de première importance. En revanche, pour le suivi de la dynamique des forêts naturelles, il est nécessaire, notamment, de mettre en place des parcelles permanentes adaptées à la richesse spécifique des milieux tropicaux humides qui permettent de suivre la croissance des individus, le taux de mortalité ou encore le nombre de juvéniles recrutés chaque année dans le peuplement. Ces exemples illustrent le fait (i) que les processus étudiés sont de nature différente, continu pour la croissance, discret pour les autres, (ii) que les arbres partagent un environnement commun qu'il soit physique ou génétique et enfin (iii) que les observations sont réalisées pour chaque individu sur plusieurs années (données longitudinales). À l'heure actuelle où la compréhension globale des processus biologiques apparaît de plus en plus réaliste, grâce notamment aux capacités modernes d'acquisition de données (phénotypage, génotypage haut-débit, données satellites, etc), la modélisation mathématique s'avère une étape clef.

Mais modéliser dans son ensemble un système complexe caractérisé par des variables de natures différentes en tenant compte des dépendances, intrinsèques ou induites par des proximités spatiales ou temporelles est un défi que j'ai essayé de relever selon différentes approches ou cadres de modelisation (fréquentiste ou bayesien) et en recourant à l'utilisation d'outils variés. Deux points de vues sont possibles pour présenter mes recherches. Je pourrais prendre celui du statisticien qui, décrivant ses méthodes, met en évidence leur « grande généralité » en soulignant par la suite l'ensemble, évidemment très large, des applications possibles. Mais, cela déformerait la réalité ou du moins la façon dont j'ai abordé ma recherche. Ce sont bien les questions biologiques qui ont motivé mes recherches en statistique. Il est donc naturel que ce soit au travers des applications que je présenterai mes activités de recherches en bio-statistiques.

1. l'héritabilité correspond à la part de variance phénotypique relevant de la variance génotypique

1

Contexte écologique et cadre statistique

Ecouter la forêt qui pousse plutôt que l'arbre qui tombe
F. Hegel (1770-1831)

1.1 Les forêts tropicales

1.1.1 Contextes

Le devenir des forêts est désormais l'une des préoccupations majeures du 20$^{\text{ième}}$ siècle : 2011 a été proclamée année internationale des forêts par l'assemblée générale des Nations Unies qui n'a pas manqué de souligner la nécessité d'une gestion durable de tous les types de forêts et d'affirmer la nécessité d'efforts concertés de sensibilisation à tous les niveaux pour renforcer la conservation et le développement viable de tous les types de forêts dans l'intérêt des générations présentes et futures (résolution 61/193). La définition de la gestion durable des forêts proposée lors de la conférence ministérielle sur la protection des forêts en Europe (FAO, 2010) est la suivante : *« la gestion durable des forêts signifie la gestion et l'utilisation des forêts et des terrains boisés d'une manière et à une intensité telle qu'elles maintiennent leur diversité biologique, leur productivité, leur capacité de régénération, leur vitalité et leur capacité à satisfaire, actuellement et pour le futur, les fonctions éco-*

logiques, économiques et sociales pertinentes aux niveaux local, national et mondial, et qu'elles ne causent pas de préjudices à d'autres écosystèmes ».

Ces préoccupations sont justifiées par l'importance que revêtent les forêts pour de multiples acteurs et à de multiples échelles. À l'échelle locale, elles constituent une ressource vitale pour les populations ; à l'échelle nationale, elles sont une source de devises grâce à l'exploitation du bois ; à l'échelle régionale et globale, elles sont un réservoir de biodiversité [1] et elles participent à la régulation du climat (Millennium Ecosystem Assessment, 2005), en jouant notamment un rôle important dans la régulation de la concentration en dioxyde de carbone (CO_2) de l'atmosphère à travers la séquestration biologique du carbone. Dans la plupart des pays tropicaux, le bois, d'énergie ou d'œuvre, provient principalement des forêts naturelles. La superficie du domaine forestier permanent naturel tropical est estimée à 761 millions d'hectares, dont un peu plus de la moitié (53 % soit 401 millions d'hectares) est affectée à la production (Blaser et al., 2011).

1.1.2 Enjeux et expérimentations

L'enjeu consiste aujourd'hui à conserver la biodiversité des forêts tropicales et à les gérer durablement, c'est-à-dire à exploiter leurs ressources en préservant à long terme leurs fonctions écologiques, économiques et sociales. L'équipe dans laquelle je suis actuellement a comme objectif de démontrer qu'il existe des conditions écologiques et socio-économiques qui permettent une utilisation raisonnée des forêts. Cette hypothèse est-elle réalisable, je ne puis le dire, mais elle reste au centre de nombreux débats internationaux (FAO, 2010). Or la nécessité de protéger et de gérer durablement un écosystème dans son ensemble conduit à le considérer non plus comme un ensemble indépendant de processus biologiques mais comme un ensemble de processus interdépendants : *le produit de la multitude d'interactions entre organismes vivants dans des milieux en changement* (Barbault and Weber, 2010). Élaborer des règles de gestion compatibles avec un renouvellement des ressources nécessite de mieux comprendre le fonctionnement à la fois de la dynamique de l'écosystème dans son ensemble mais aussi celle des biens et des services qu'il rend. Pour aborder ces questions il est impératif de disposer d'un grand nombre d'informations provenant souvent de sources variées et d'élaborer la mise en place de plans d'aménagement. Ceux-ci doivent permettre une ges-

1. Elles abritent une faune et une flore d'une grande richesse. La forêt tropicale humide, qui ne couvre que 6 % de la planète, renferme à elle seule 50 % à 80 % des espèces animales et végétales terrestres : 80 % des insectes, 84 % des reptiles, 91 % des amphibiens, 90 % des primates, 70 % des espèces végétales connues dont près de 50 000 espèces d'arbres. On estime que plus de 1.5 milliards de personnes en vivent directement ou indirectement

tion durable des forêts tropicales. Mais la définition de ces plans nécessite une connaissance préalable des forêts. Or celles-ci sont trop vastes pour être inventoriées dans leur totalité et présentent de fortes hétérogénéités spatiales et cela à différentes échelles. Les grands facteurs environnementaux (pluviométrie, topographie, substrat géologique) tout comme les perturbations, et notamment anthropiques, façonnent leur structure et leur composition floristique. Au niveau d'une concession, une zone de forêt marécageuse ou une zone de forêt très perturbée (« secondarisée ») ne présenteront pas les mêmes caractéristiques en termes de dynamique, de valeurs économiques ou écologiques qu'une zone de forêt non perturbée sur des sols non contraints. Pour orienter les gestionnaires dans leurs prises de décisions, les inventaires forestiers sont nécessaires. En plus de l'espèce et du diamètre de chacun des arbres répertoriés, ceux-ci offrent la possibilité de quantifier d'autres caractéristiques importantes de la forêt : hauteur du peuplement, type de sol, végétation herbacée, etc. Pour autant, les inventaires forestiers ne sont généralement pas suffisants pour comprendre (ou encore prédire) à long terme l'ensemble des différents services de la forêt et en particulier, la production de bois ou la biodiversité. Les parcelles permanentes, qui permettent de suivre sur la durée des zones spécifiques, s'avèrent aussi cruciales pour une gestion à long terme des forêts tropicales. L'étude de la dynamique forestière nécessite souvent des analyses à une échelle de temps et d'espace supérieure à celle qu'il est possible d'observer sur le terrain ; c'est pourquoi il est nécessaire d'avoir recours à la modélisation (Lourmas, 2003). Bien que de nombreux modèles aient été développés, les tentatives effectuées en matière d'aménagement et de gestion des forêts tropicales se heurtent, entre autres, à une compréhension insuffisante des phénomènes qui régissent la dynamique des populations des espèces d'arbres qui les constituent. Plusieurs raisons peuvent être invoquées, parmi lesquelles

1. la richesse de ces forêts, plus de 300 espèces à l'hectare,
2. la complexité des processus étudiés et leur interaction avec l'environnement biotique ou abiotique,
3. la longévité de ces écosystèmes,
4. l'échantillonnage qui doit être adapté aux enjeux spécifiques.

Nous avons la chance au CIRAD d'avoir des dispositifs expérimentaux d'une rare richesse. Le dispositif de M'Baïki en Afrique centrale (Bedel et al., 1998) et celui de Paracou en Guyane française sont historiquement les plus anciens. Le premier, mis en place en 1982 avec la collaboration de l'état centrafricain (cf figure 1.1), et le second installé en 1984 (cf figure 1.2), ont été définis pour quantifier l'impact de traitements sylvicoles sur la dynamique de régénération des essences forestières commerciales. Par la suite, dès 1992, toutes les

espèces et tous les arbres ont été mesurés. Les deux expérimentations sont globalement similaires (cf figures 1.1 et 1.2) et représentent chacune une surface échantillonnée d'environ 40 hectares divisés en 10 parcelles de 4 hectares.

Les parcelles, choisies aléatoirement, ont subi ou non un traitement sylvicole : tous les arbres commerciaux d'un diamètre supérieur à 80 cm ont été exploités (traitement 1) et certaines de ces parcelles ont subi, deux ans plus tard, un traitement supplémentaire qui consistait à « empoisonner » tous les arbres non commerciaux de plus de 50 cm (traitement 2). Les autres parcelles ont été laissées en l'état et constituaient les témoins (traitement 0). Depuis l'installation des parcelles, tous les arbres dont le diamètre à hauteur de poitrine est supérieur à 10 cm (« diameter at breast height », DBH) sont mesurés (annuellement à M'Baïki, tous les deux ans à Paracou). Les arbres qui meurent entre deux temps de mesure sont répertoriés ainsi que tous les individus qui atteignent le diamètre de 10 cm. Ainsi nous disposons, par exemple à M'Baïki, d'information sur

- 239 taxons dont 191 ont été déterminés au niveau de l'espèce, 36 ont été identifiés comme morpho-espèces et restent indéterminés, 10 ont été déterminés au niveau du genre et 2 ont été identifiés comme différents (noms vernaculaires différents) mais ont le même nom botanique. En 2012, nous disposions de plus de 200,000 points de mesures individuelles pour la croissance et la mortalité et de plus de 100,000 pour le recrutement.
- les traits éco-physiologiques des espèces : la guilde de régénération, la phénologie de la feuillaison, la densité du bois, le taux de croissance maximal et le diamètre maximal
- les types de sols dont la cartographie a été réalisée en 1992 (Ceccato et al., 1992)

Parallèlement, grâce aux nombreux projets de recherche obtenus et gérés avec et par les collègues de l'équipe, nous disposons d'autres sources de données. On peut citer par exemple

1. pour les forêts du bassin du Congo, la base de données CoForChange [2] et CoForTips [3]. Cette base de données spatiales rassemble des informations sur l'abondance des espèces dans les inventaires de 11 concessions forestières, la géologie, les sols et la topographie, le climat (précipitations annuelles, nombre de mois secs, date de démarrage et durée de la saison sèche), la localisation des villes, villages, routes et pistes mais aussi des informations provenant de l'analyse d'images Spot, Mo-

2. http://www.coforchange.eu/fr/
3. http://www.cofortips.org/

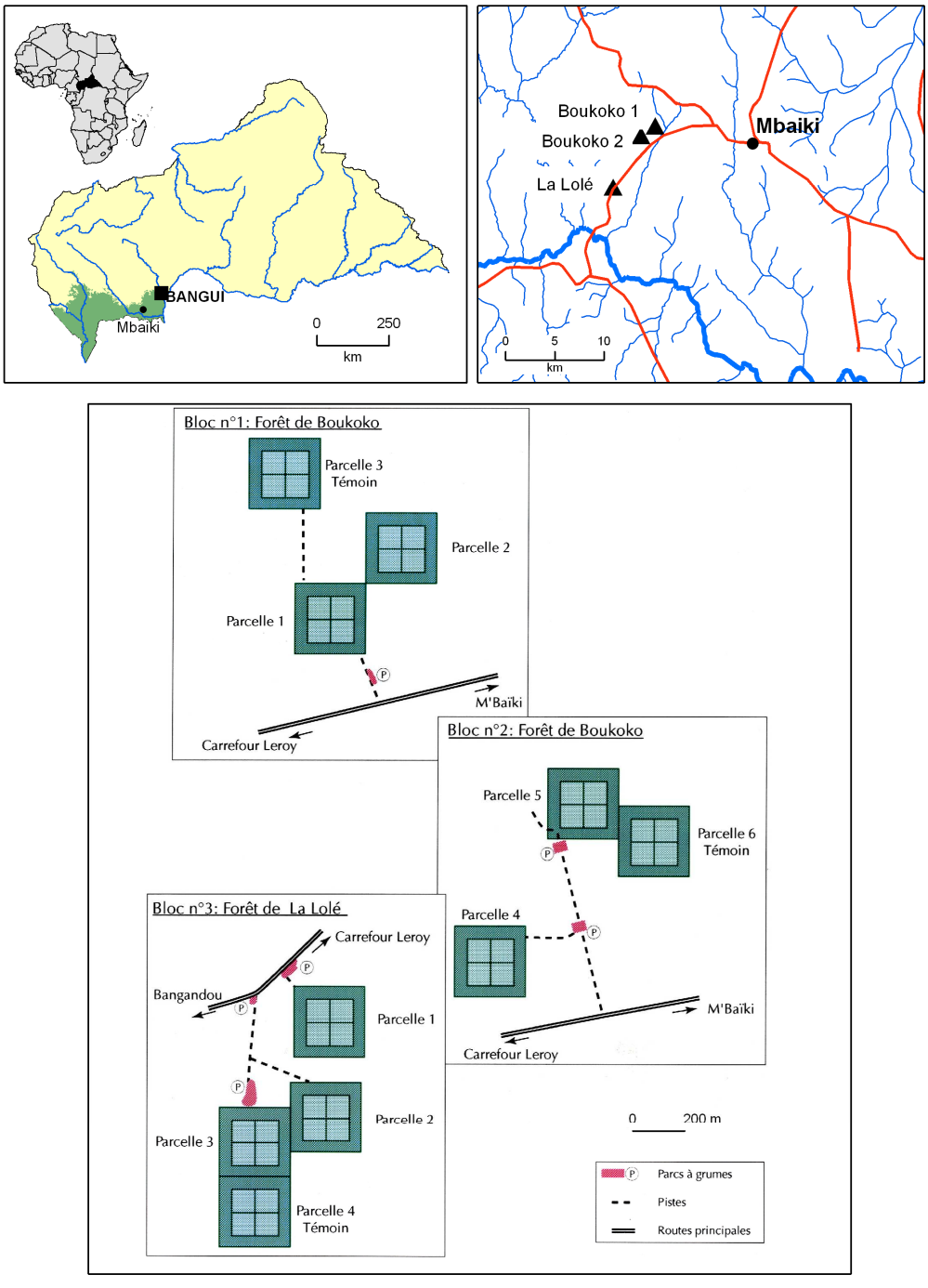


FIGURE 1.1: Dispositif expérimental de M'Baïki installé en 1982

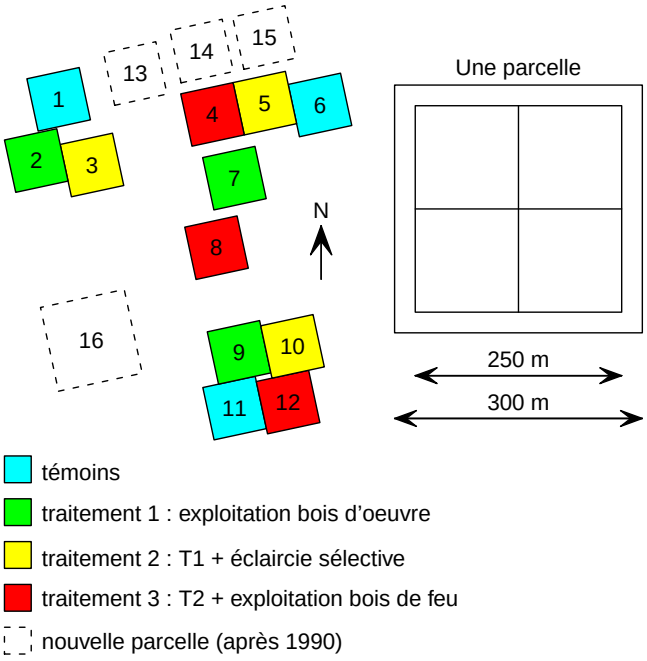


FIGURE 1.2: Dispositif expérimental de Paracou installé en 1984

dis et Landsat. Les données couvrent cinq pays majeurs de la zone : le Cameroun, la Centrafrique, le Congo, le Gabon et une partie de la République démocratique du Congo. La surface inventoriée représente plus de 60,000 hectares.

2. pour les forêts amazoniennes, du Bassin du Congo et sud-asiatiques, les données mises en partage par les partenaires du projet TmFo [4], concernent plus de 24 sites expérimentaux suivis annuellement depuis des époques variables et distribués sur les 3 continents avec un total d'environ 500 parcelles permanentes.

Ainsi, les enjeux sont nombreux et ambitieux. Mais les données dont on dispose permettent de concevoir des éléments de réponses et apporter des premières explications.

1.2 Les plantations

1.2.1 Contexte

Il est peut-être surprenant d'évoquer les plantations forestières au sein même de zones où les forêts naturelles semblent en abondance. Pourtant, les plantations s'avèrent sûrement l'une des solutions pour répondre à la demande toujours plus forte en bois et pour préserver des espaces naturels. Les forêts naturelles et les plantations d'eucalyptus forment les deux filières du bassin d'approvisionnement urbain en bois-énergie de Pointe- Noire, capitale économique du Congo (Nkoua and Gazull, 2013). L'évolution des attentes dans le domaine de la foresterie, qu'il s'agisse de la conservation de la biodiversité, de la séquestration du carbone, de la certification, de la restauration écologique mais aussi de l'évolution des usages, impose de reconsidérer les concepts autour desquels sont bâties les plantations forestières (Marien and Mallet, 2004). Autant, il y a encore quelques années, l'objectif des plantations était de répondre à une demande massive de l'industrie et en particulier celle de la pâte à papier, autant aujourd'hui, le bois doit être utile à différents acteurs simultanément : bois énergie, construction, puits de carbone. Il est donc nécessaire de reconsidérer non seulement les concepts autour desquels sont bâties les plantations forestières mais aussi les schémas d'amélioration.

Les plantations forestières couvrent actuellement environ 200 millions d'hectares alors qu'elles ne représentaient que 30 millions d'hectares en 1970 (FAO, 2010). Cette forte augmentation est liée principalement à la réduction de l'exploitation des forêts naturelles, à l'accroissement mondial de la

4. http://tmfo.org/

consommation en bois de chauffe (due à l'augmentation de la population) et à la demande en pâte à papier (FAO, 2010). La production de papier est passée de 75 millions de tonnes en 1961 à environ 350 millions en 2005. Parmi les espèces utilisées en plantations industrielles, le genre eucalyptus est l'essence forestière feuillue la plus plantée au monde (FAO, 2010). Depuis la fin du $XIX^{\text{ième}}$ et le début du $XX^{\text{ième}}$ siècle, l'intérêt porté au genre eucalyptus n'a cessé de croître (Vigneron et al., 2000). En raison de la grande variabilité des espèces - plus de 700 composent le genre eucalyptus - celui-ci est présent dans de nombreuses zones géographiques. En 2002, la FAO recensait plus de 90 pays utilisateurs de ce genre. Les plantations se situent pour l'essentiel en zone tropicale et subtropicale, mais s'étendent aussi aux régions tempérées chaudes méditerranéennes (Portugal, Espagne en particulier). Ce genre est essentiellement utilisé comme bois de chauffe ou pour la fabrication de pâte à papier : son rendement papetier [5] est très supérieur à celui des autres feuillus. Il est également employé comme bois d'œuvre, ou pour ses huiles essentielles en industrie pharmaceutique et cosmétique.

Au Congo, un programme d'amélioration d'eucalyptus a été mis en place dans les années 70. Les recherches en amélioration génétique de l'eucalyptus ont accompagné et favorisé le développement de plantations industrielles dédiées à la production de bois pour les industries papetières. Les programmes d'amélioration ont pour objectif de sélectionner les meilleurs individus (génotypes) d'une population pour engendrer les générations suivantes. Ils visent à optimiser les valeurs d'un ou plusieurs caractères phénotypiques en utilisant la variabilité génétique présente au sein des espèces. Les "améliorateurs" se basent sur la valeur phénotypique pour estimer la valeur génétique et ainsi sélectionner les individus qui serviront de géniteurs pour les générations suivantes (Lynch and Walsh, 1998). Au Congo, initialement, des hybrides naturels ont été utilisés pour les plantations. Mais, ces plantations ont vu leur production très rapidement plafonner (Vigneron, 1991). Pour résoudre ce problème, plusieurs genres d'eucalyptus et différents programmes d'amélioration ont été envisagés. Finalement, le schéma de sélection récurrente réciproque (SRR) et deux espèces d'eucalyptus, *urophylla* et *grandis* ont été choisis. La première espèce d'eucalyptus présente de bonnes capacités d'adaptation aux conditions climatiques du Congo tandis que la seconde est connue pour son importante potentialité de croissance (Vigneron, 1991). Le principe de la SRR est d'améliorer conjointement deux groupes d'individus, de manière à obtenir des hybrides concentrant certains caractères spécifiques à chacun des deux groupes (parentaux). Dans ce cadre, le caractère cible est la croissance avec, en ligne de mire, un objectif clairement affiché : la sélection précoce

5. ratio quantité de bois utilisé sur quantité de pâte produite

des meilleurs génotypes. Mais les corrélations juvéniles/adultes sont relativement faibles dans le cas de l'eucalyptus (Bartholomé et al., 2013). De plus, les schémas d'amélioration reposent sur des dispositifs expérimentaux lourds à mettre en place qui couvrent plusieurs dizaines d'hectare.

1.2.2 Enjeux et expérimentations

L'arrivée de méthodes de génotypage et de la génétique moléculaire ont modifié les approches et ouvert la voie aux méthodes d'amélioration assistée par marqueurs. L'objectif est d'accroître l'efficacité de la sélection par unité de temps. Ces méthodes ont considérablement modifié la façon de concevoir les plans d'expérience, de collecter l'information phénotypique ou moléculaire par l'utilisation de techniques à haut débit. Ces méthodes permettent d'acquérir en un temps relativement court une masse importante d'information génétique (« single-nucleotide polymorphism » , données d'expression) ainsi qu'un ensemble plus large de caractères d'intérêt. Ces techniques offrent de nouveaux outils pour répondre aux nouveaux défis des plantations mais soulèvent aussi de nombreuses nouvelles questions, en particulier celle de la sélection génomique (Meuwissen et al., 2001).

Le CIRAD dispose de sites expérimentaux pour répondre à ces nouveaux enjeux, notamment celui du CR2PI à Pointe-Noire. De nombreux tests y ont été mis en place depuis le début des années 1990 : test de provenance, clonaux ou de descendance. On peut aussi citer une expérimentation nouvelle qui a été mise en place dans le cadre du projet Abiogen porté par Jean-Marc Gion qui consiste à suivre de manière journalière la croissance des arbres issus d'une famille et les conditions environnementales d'ensoleillement et d'humidité (Bartholomé et al., 2013).

1.3 Enjeux de modélisation

Analyser les données issues de ces écosystèmes nécessite certaines précautions et des méthodes d'analyses adéquates doivent être employées. En particulier celles-ci doivent tenir compte de certaines caractéristiques de ces écosystèmes.

- La corrélation temporelle : l'étude de la dynamique d'un écosystème se fonde sur le suivi au cours du temps des individus qui le composent. Il est donc impératif de définir le bon niveau de description, de l'individu à l'écosystème, en tenant compte des dépendances temporelles.
- La corrélation spatiale : l'auto-corrélation est commune dans les données écologiques (Legendre, 1993). Elle traduit le fait que la plupart des

processus biologiques sont contagieux : leurs effets se manifestent sur des surfaces continues à des échelles variées (ex. : croissance, dispersion, attaque de pathogènes). En conséquence, des observations proches spatialement ont tendance à être comparables (autocorrélation positive).
- La richesse spécifique : les forêts tropicales humides sont composées, selon les régions du monde, par un ensemble d'espèces dont le nombre peut varier entre 300 à 500 par hectare. Cela se traduit par des compétions inter-spécifiques souvent fortes (corrélation négative) dont la modélisation reste problématique. De plus, l'abondance des espèces peut être très variable et il est fréquent de n'observer qu'un nombre restreint d'individus par espèce. Il est alors illusoire de vouloir construire des modèles spécifiques.

Les principaux enjeux sont de comprendre et prédire les dynamiques des écosystèmes forestiers sur la base d'informations locales et souvent lapidaires. Ces défis sont explicités dans l'article de Clark (2005) :

- Comment combiner différentes sources d'informations ?
- Comment intégrer la connaissance d'un processus à une échelle locale pour l'étendre à une échelle globale ?
- Comment prendre en compte les différentes sources d'incertitudes dans les prédictions ?

1.4 Cadres méthodologiques

Cette section n'a pas pour vocation de présenter en détail les cadres méthodologiques ; de nombreux ouvrages sont consacrés à ces domaines de recherches. L'objectif est avant tout de faire une rapide présentation pour montrer d'une part comment ces modèles sont utilisés en science de l'environnement et en particulier en écologie et d'autre par pour présenter les questions méthodologiques auxquelles j'ai tenté de répondre.

1.4.1 Les modèles hiérarchiques bayésiens

Avec le développement des méthodes et une puissance de calcul devenue non limitante, les approches bayésiennes se sont largement développées en écologie statistique[6]. En particulier les modèles hiérarchiques bayésiens (MHB) ont connu un fort engouement dès le début des années 2000 (Gimenez et al., 2014). Les MHB sont conçus comme une succession de sous-modèles dans lesquels les paramètres d'un niveau donné dépendent de ceux du niveau

6. Le groupe « Environnement » de la SFDS m'a offert l'occasion en 2011 de faire une revue bibliographique sur ce sujet (Mortier et al., 2011)

suivant (Wikle, 2003). De manière formelle, un MHB se décompose en au moins trois niveaux :

1. le « *data level* » qui consiste à modéliser la distribution des observations, x, conditionnellement à un processus d'intérêt ϑ : $[x|\vartheta]$,
2. le « *process level* » qui consiste à modéliser la distribution du processus : $[\vartheta|\varphi]$, φ étant un vecteur de paramètres associés à ϑ,
3. le « *parameter level* » qui consiste à stipuler les lois a priori de φ

où $[x]$ dénote la distribution de x tandis que $[x|\vartheta]$ dénote la distribution de x sachant ϑ (Cressie et al., 2009). L'intérêt de l'approche hiérarchique est d'être extensible à plusieurs niveaux d'hypothèse. Ainsi, il est tout à fait envisageable que ϑ soit lui même dépendant d'un autre processus caché. De plus, il est souvent possible de représenter ces modèles sous forme graphique en utilisant un graphe acyclique orienté (« Directed Acyclic Graph », DAG), ce qui facilite la compréhension du modèle et sa construction (cf figure : 1.3).

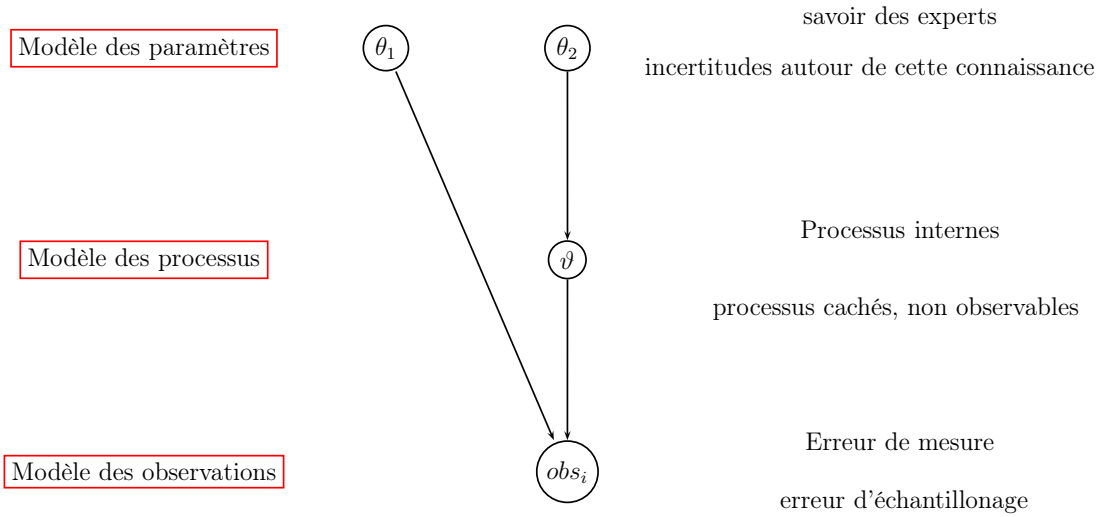


FIGURE 1.3: Présentation graphique d'un modèle hiérarchique bayésien

L'intérêt pour les MHB est dû à leurs flexibilités en permettant de décomposer la complexité des phénomènes biologiques en une série de sous-modèles plus simples (Banerjee et al., 2004; Parent and Bernier, 2007). Les hypothèses classiques d'indépendance sont remplacées par des hypothèses d'indépendance conditionnelle. Enfin, le cadre bayésien offre l'avantage d'incorporer des connaissances *a priori* sous diverses formes, ce qui est particulièrement adapté à la recherche en biologie. Les MHB sont devenus d'autant plus populaires qu'ils s'appuient sur des méthodes d'inférence accessibles à des non statisticiens grâce à des logiciels tels que WinBugs (Lunn et al., 2000), JAGS (Plummer, 2003). De plus, dès le début des années 2000, une série impressionnante d'articles a contribué à les populariser. En 2003, Wikle publie dans *International statistical review* un premier article intitulé « *Hierarchical Models in Environmental Science* ». L'auteur met en avant la complexité des écosystèmes biologiques ainsi que la nécessité de considérer les processus simultanément pour présenter l'approche hiérarchique et en particulier bayésienne comme une solution élégante et efficace. Cet article est suivi en 2004 par celui de Ellison dans *Ecology Letters* qui présente plus en détail les méthodes d'inférence bayésienne pour l'écologie (Ellison, 2004). L'article de Clark en 2005 intitulé « *Pourquoi les sciences de l'environnement deviennent-elles bayésiennes* » (« *Why environmental scientists are becoming Bayesians* ») marque un aboutissement. Par la suite, deux autres articles McMahon and Diez (2007) et Cressie et al. (2009) viendront compléter la liste.

La souplesse de la modélisation hiérarchique bayésienne combinée aux outils informatiques disponibles (et gratuits) permet désormais d'envisager des modèles de plus en plus « réalistes » pour aborder des questions de plus en plus complexes. Mais cette approche peut aussi aboutir à une complexification excessive et risquée (Gimenez et al., 2014). Celle-ci peut donc s'avérer utile et féconde mais peut aussi conduire à des problèmes d'inférence et de stabilité, voir à des résultats erronés. Il existe de nombreux outils qui permettent de quantifier la qualité des modèles ou de les comparer. Le critère de Bayes est le plus classique. En pratique, il est quasiment impossible de le calculer explicitement. Des versions approchées ont été proposées, notamment en se basant sur la moyenne harmonique obtenue sur les échantillons des chaînes de Markov de la vraisemblance *a posteriori* (Raftery et al., 2007). Mais les résultats restent instables en particulier dans les modèles complexes. D'autres critères ont alors été proposés tel que le DIC, *deviance information criterion* (Spiegelhalter et al., 2002), les versions bayésiennes de l'AIC ou BIC basées sur le facteur de Bayes que sont l'AICM et BICM (Raftery et al., 2007) ou encore le *posterior predictive p-value* (Gelman et al., 1996) et le *posterior predictive loss* (Gelfand and Ghosh, 1998). Le premier ensemble de critères (DIC, AICM ou BICM) reflète la qualité d'ajustement du modèle tandis que

le second la qualité de prédiction (Guisan and Zimmermann, 2000; Banerjee et al., 2004). Mais en pratique, ces outils restent encore peu employés. Il y a un besoin évident de mettre en place des stratégies d'évaluation des nouveaux modèles pour déterminer les bénéfices que l'on peut en retirer.

Mes contributions J'ai employé ce cadre statistique dès mon arrivée au CIRAD pour développer des modèles qui tiennent compte par exemple des corrélations spatiales et temporelles dans le contexte de l'amélioration génétique,de la dynamique de données ordinales, des champs spatiaux non-gaussiens. Cela s'est concrétisé directement par l'encadrement de quatre étudiants de « master », le co-encadrement de trois étudiants en thèse et un en post-doctorat. L'utilisation de ce cadre statistique m'a permis de publier six articles méthodologiques (Guillot et al., 2005a; Chaubert et al., 2008; Flores et al., 2009; Chagneau et al., 2011; Garreta et al., 2012; Mortier et al., 2013) et un article appliqué (Guillot et al., 2005b).

1.4.2 Les modèles spatiaux

Élaborer des modèles qui permettent d'extrapoler des résultats obtenus à partir d'un échantillonnage effectué sur une zone donnée à l'ensemble d'un écosystème reste un enjeu majeur. Les outils de la *statistique spatiale* sont maintenant incontournables (Cressie, 1991; Banerjee et al., 2004). Développée par Matheron (1963) à partir des travaux de Krige (1951), le krigeage est une méthode stochastique d'interpolation spatiale. Celle-ci permet de prédire la valeur d'une variable en des sites non échantillonnés. La prédiction est une combinaison linéaire sans biais et de variance minimale (Baillargeon, 2005). L'une des particularités de ces écosystèmes réside dans le nombre et la nature variée des observations : des données de présence/absence ou d'abondance pour les espèces mesurées sur des parcelles d'inventaires ; des données environnementales comme la pluviométrie, l'altitude ou encore la couleur du sol échantillonnées de manière aléatoire et en un nombre limité de sites. La modélisation des problèmes géostatistiques proposée par Diggle et al. (1998), appelée *model-based geostatistics*, offre une méthode unifiée pour traiter des réponses qui peuvent être des présence/absence, qui peuvent être issues de comptages (abondance) ou encore continues (Banerjee et al., 2004; Christensen and Waagepetersen, 2002). La démarche est la suivante. On suppose que le modèle dont sont issues les données vérifie les hypothèses suivantes. On suppose tout d'abord qu'il existe un processus stationnaire S, d'espérance nulle ($\mathbb{E}[S(\mathbf{s})] = 0$) et covariance égale à $\mathrm{Cov}[S(\mathbf{s}), S(\mathbf{s} + \mathbf{h})] = \sigma^2\rho(\mathbf{h})$. Dans un deuxième temps, on suppose que conditionnellement au processus S, les variables aléatoires Y_i, $i = 1, \ldots, n$ sont mutuellement

indépendantes de densité $f_i(y|S_i) \equiv f(y; M_i)$, densité qui ne dépend que des valeurs des espérances conditionnelles $M_i = \mathbb{E}(Y_i|S_i)$. De manière similaire à la démarche employée dans le cadre des modèles linéaires généralisés, le prédicteur linéaire η_i est alors linéairement expliqué au travers d'une fonction de lien g par un ensemble de covariables x_i de sorte que :

$$\eta_i = g(M_i) = S_i + x_i'\beta.$$

β désigne les paramètres inconnus. Avec ces hypothèses, le prédicteur de $S(\mathbf{s})$, appelé prédicteur linéaire généralisé, est défini par $S^*(\mathbf{s}) = \mathbb{E}[S(\mathbf{s})|Y]$. Cette construction est donc très similaire à celle de la modélisation hiérarchique présentée précédemment (voir section 1.4.1) ou à celle des modèles linéaires généralisés à effets aléatoires (McCulloch et al., 2008). Les estimations des paramètres de covariance du processus spatial sont alors obtenues soit par maximum de vraisemblance ou maximum de vraisemblance restreinte.

Le problème lié à la modélisation multivariée repose sur l'explicitation de la structure de dépendance intra et inter processus. Une façon de modéliser la dépendance entre les variables est de construire la matrice de covariance du vecteur $\mathbf{Y} = (Y_1, \ldots, Y_K)$ où chaque $Y_k, k = 1, \ldots, K$ est un processus spatial. C'est sous cette forme que l'on modélise la dépendance dans le cas du krigeage et du cokrigeage. Cette matrice se décompose en différents blocs de la forme $\mathrm{Cov}[\mathbf{Y}(\mathbf{s}_i), \mathbf{Y}(\mathbf{s}_j)]$. Ces blocs sont des matrices de dimension $K \times K$ qui ne sont pas forcément symétriques. En revanche, la matrice de covariance de $\mathbf{Y}$, de dimension $Kn \times Kn$, doit être définie positive quels que soient le nombre et le choix des points échantillonnés. La difficulté est de proposer des modèles qui définissent des matrices de covariance valides. Les modèles classiques sont ceux dits à covariance proportionnelle (ou modèle de corrélation intrinsèque) (Wackernagel, 2003) ou les modèles linéaires de corégionalisation (Grzebyk and Wackernagel, 1994; Wackernagel, 2003). Ces modèles de covariance sont valides, mais peu flexibles car ils sont basés sur un nombre restreint de fonctions élémentaires. Barry and Ver Hoef (1996) définissent une nouvelle famille de variogrammes valides basés sur des fonctions de carré intégrable, dites « fonctions moyennes mobiles » (Barry and Ver Hoef, 1996; Ver Hoef and Barry, 1998). Ce modèle de covariance décrit la structure de dépendance existant entre des processus aléatoires spatiaux Z_k, $k = 1, 2, \ldots, K$ construits par intégration du produit de convolution d'une fonction moyenne mobile f_k et d'un mélange de bruits blancs. Cette méthode, outre le fait qu'elle définisse par construction des fonctions de covariance ayant de bonnes propriétés, présente l'avantage d'être flexible, de pouvoir générer des processus non-stationnaires et anisotropes ou encore des processus non-gaussiens, par exemple par la convolution d'une fonction moyenne mobile avec un processus de Poisson ou Gamma (Gaetan and Guyon, 2008). Cependant, comme

dans tout problème statistique, il convient de bien choisir le processus et les fonctions moyennes mobiles.

Mes contributions La prise en compte des dépendances spatiales est une nécessité en écologie forestière ou pour modéliser les compétitions entre les individus au sein de plantations. Selon le contexte biologique et l'échantillonnage, j'ai été amené à utiliser ou à développer des modèles spatiaux variés.

- Des modèles « conditional autoregressive » (CAR) ou « simultaneous auto-regressive » (SAR) dans le cadre où l'échantillonage était sur treillis (lattice, en anglais),
- des modèles de géo-statistiques lorsque le domaine spatial était continu.

L'utilisation des outils de la statistique spatiale m'a permis de publier sept articles méthodologiques (Guillot et al., 2005a; Picard et al., 2008a, 2009; Flores et al., 2009; Chagneau et al., 2009, 2011; Garreta et al., 2012) et trois appliqués (Guillot et al., 2005b; Picard et al., 2008a; Fayolle et al., 2012).

1.4.3 Les modèles de mélanges

L'une des spécificités des écosystèmes forestiers et des milieux tropicaux en général, réside dans leurs grandes richesses biologiques. Celle-ci a conduit de nombreux auteurs à s'interroger sur les mécanismes de co-existences qui permettent une telle richesse (Hutchinson, 1961; Hubbell, 2001). Mais là n'est pas mon propos. L'une des approches historiquement utilisée pour gérer cette biodiversité repose sur la construction de groupes. Ceux-ci ont été définis soit sur le tempérament des espèces (Swaine and Whitmore, 1988), soit sur les caractéristiques fonctionnelles (Steneck and Dethier, 1994) ou encore sur les propriétés éco-morphologiques des espèces (Bellwood and Wainwright, 2001). Mais ces outils de construction *a priori* de groupes ne sont pas adaptés à la principale question qui concerne la prédiction (Dunstan et al., 2011, 2013; Hui et al., 2013). J'ai donc choisi d'orienter mes recherches sur la construction de groupes, « modèles centrés », adaptés à la prédiction. Le cadre méthodologique que j'ai choisi se fonde sur les modèles de mélanges. Les approches par mélange ont été largement utilisées pour l'estimation paramètrique de distribution de variables aléatoires en les modélisant par exemple comme une somme de plusieurs gaussiennes (appelées noyaux), en génétique où ils ont été utilisés pour l'analyse de données d'expression et en génomique (Daudin et al., 2008; Tadesse et al., 2005). En écologie, les modèles de mélange sont, il me semble, encore relativement peu usités comparé à d'autres types de modèles tels que les modèles à effets aléatoires (McCulloch et al., 2008). Ils sont principalement employés pour modéliser l'abondance d'un grand nombre

d'espèces, ou *archetypes*, simultanément (Dunstan et al., 2011, 2013; Hui et al., 2013)

De manière générale, les modèles de mélange se basent sur l'hypothèse que les observations $x_i, i = 1, \ldots, n$ sont issues d'un mélange de distributions (McLachlan and Peel, 2004) :

$$f(x_i) = \sum_{k=1}^{K} \pi_k f_k(x_i, \theta_k), \quad \text{où} \quad \sum_{k=1}^{K} \pi_k = 1$$

où K est le nombre de composantes du mélange, π_k les poids du mélange, f_k une loi de probabilité et θ_k les paramètres de cette loi. Les modèles de mélanges communément utilisés sont les mélanges gaussiens, qui supposent que f_k est une loi gaussienne et $\theta_k = (\mu_k, \sigma_k^2)$. L'enjeu, en plus de ceux qui concernent l'estimation et l'assignation des observations aux groupes, reste celui du choix du nombre de groupes. Dans le contexte fréquentiste, ce choix s'effectue soit en utilisant des critères classiques tels que l'AIC (Akaike, 1974) ou le BIC (Schwarz, 1978) où le nombre de paramètres est calculé en tenant compte, outre du nombre de paramètres de la distribution considérée, du nombre de proportions du mélange moins un. Plus récemment Biernacki et al. (2000) ont proposé l'« *integrated completed likelihood* » (ICL) qui s'interprète comme le critère BIC auquel est ajouté une pénalisation qui tient compte de la qualité de la classification. Dans le cadre bayésien, différentes méthodes existent pour estimer le nombre de groupes. Notamment, il est possible de supposer le nombre de groupes comme une variable aléatoire et d'utiliser des algorithmes MCMC à sauts réversibles (RJ-MCMC) (Richardson and Green, 1997). Mais la mise en œuvre pratique de cet algorithme est périlleuse en particulier si l'espace des paramètres est important. Néanmoins, les modèles de mélange fournissent plusieurs avantages par rapport aux approches heuristiques basées sur des critères métriques tels que les k-means, notamment un cadre formel pour incorporer des variables explicatives. Les modèles de régression en mélange (McLachlan and Peel, 2004) considèrent que la loi de y est un mélange de régression linéaire généralisée dont chaque modèle est gouverné par différents paramètres. Soit g la fonction de lien canonique associée à la fonction de distribution supposée appartenir à la famille exponentielle. Le modèle s'exprime de la façon suivante :

$$\begin{aligned} f(y_i) &= \sum_{k=1}^{K} \pi_k f(y, \theta_{ik}) \\ g(\theta_{ik}) &= x_i' \beta_k \end{aligned}$$

où x est un ensemble de covariables et β_k un vecteur de paramètres indexés par le groupe k. Ce cadre mathématique permet simplement d'abor-

der la question du déterminisme environnemental dans un contexte multi-spécifiques.

Mes contributions L'utilisation que j'ai pu faire des modèle de mélange est plus récente. Néanmoins, j'oriente actuellement très largement mes recherches et mes collaborations sur des problématiques multi-spécifique et ce cade méthodologique me semble hautement pertinent. L'utilisation des modèles de mélange m'a déjà permis de publier deux articles méthodologiques (Mortier et al., 2013, 2015) et un article appliqué (Ouédraogo et al., 2013). De plus cela m'a permis de contribuer au « package » **flexmix** (Leisch, 2004; Grün and Leisch, 2007, 2008).

1.4.4 Réduction de dimension

L'un des enjeux actuels réside dans la capacité de choisir ou de combiner les facteurs biotiques ou abiotiques pour prédire la distribution des espèces, la dynamiques des écosystèmes ou la croissance des individus. Deux stratégies sont envisageables pour dénouer ces liens entre réponses et variables explicatives, selon que l'on cherche avant tout à expliquer un phénomène ou à le prédire. La première repose principalement sur des approches telles que la « sélection de variables » tandis que la seconde se fonde davantage sur des approches de types « compression d'information », la plus classique étant la régression sur composantes principales.

Approches par sélection de variables : les méthodes *stepwise* sont largement répandues et utilisées. Néanmoins, de nouvelles techniques de sélection plus adaptées aux données actuelles, en particulier à leur abondance, ont vu le jour dès la fin des années 90. En particulier, celles basées sur des approches régularisées dont la méthode LASSO (*least absolute shrinkage and selection operator*) introduite par Tibshirani (1996). Celle-ci a pour objectif de chercher à maximiser la vraisemblance sous la contrainte que la norme L_1 des paramètres soit plus petite qu'une constante à choisir :

$$\arg\max_{\beta} \ell(y, \beta) \quad \text{sous la contrainte} \quad ||\beta||_1 < c$$

où $||x||_1$ est la norme L_1 de x. Ce problème revient à maximiser la vraisemblance des observations pénalisées :

$$\arg\max_{\beta} \ell(y, \beta) - \lambda||\beta||_1.$$

Le choix de λ se fait classiquement par validation-croisée. L'intérêt de ces méthodes est qu'elles permettent simultanément d'estimer les paramètres et

d'éliminer les covariables x_l non pertinentes pour l'analyse en ramenant à zéro la valeur des coefficients associés (*shrinkage*). De façon générale le problème posé peut se représenter sous la forme suivante :

$$\arg\max_{\beta} \ell(y, \beta) - pen(\beta).$$

où *pen* est une pénalisation qui dépend de β. De nombreuses pénalisations ont par la suite été proposées qui assurent à ces méthodes des propriétés d'optimalité. Parmi celles-ci on peut citer les pénalisations HARD ou SCAD (Fan and Jinchi, 2010) ou les pénalisation adaptatives (Zou, 2006). Ces approches ont néanmoins l'inconvénient d'être mal adaptées lorsque les covariables présentent de fortes collinéarités, à la différence des approches qui utilisent des pénalisations impliquant la norme L_2 telle que la méthode Ridge (Hoerl and Kennard, 1970). Pour remédier à ces difficultés, Zou and Hastie (2005) introduisent la régression appelée *elastic net regression* qui combine la méthode LASSO et Ridge. Le problème s'écrit alors :

$$\arg\max_{\beta} \ell(y, \beta) - \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2$$

ou $||x||_2$ est la norme L_2 de x.

D'un point de vue bayésien, la question de la sélection de variables est aussi un champ de recherche très actif et la littérature sur le sujet est assez conséquente (Marin and Robert, 2007; O'Hara and Sillanpää, 2009). La procédure consiste à rechercher les paramètres qui sont ou non égaux à zéro. L'approche proposée par George and McCulloch (1997) consiste à introduire une variable indicatrice γ_k qui indique si le paramètre est proche de zero ou non. Selon la façon dont est introduite cette variable indicatrice, le choix de la loi *a priori* des paramètres conduit à différentes méthodes. Initialement, l'approche développée par George and McCulloch (1997), appelée *Stochastic Search Variable Selection (SSVS)* consiste à modéliser la loi *a priori* des paramètres β_k selon que l'indicatrice associée γ_k vaut un ou zéro (« slab and spike ») :

- si $\gamma_k = 1$:
$$[\beta_k|\gamma_k = 1] = \mathcal{N}(0; \tau)$$
- si $\gamma_k = 0$:
$$[\beta_k|\gamma_k = 0] = \mathcal{N}(0; c\tau)$$
où c est une constante "petite" qu'il faut régler "manuellement" et qui permet que la loi soit "piquée" autour de zéro.

Ainsi dans cette approche, la loi des paramètres est une loi de mélange :

$$[\beta_k|\gamma_k] = \gamma_k \mathcal{N}(0; \tau) + (1 - \gamma_k)\mathcal{N}(0; c\tau)$$

D'autres alternatives ont été proposées comme celles de Kuo and Mallick (1998) ou de Dellaportas et al. (2000). Enfin, d'autres approches ont été développées récemment qui ne reposent pas sur l'utilisation de la variable indicatrice γ. Parmi celles-ci on peut citer les versions bayésiennes du LASSO (Park and Casella, 2008) et de l'elastic-net (Li and Lin, 2010).

Les approches par composantes : de façon générale, l'ensemble des méthodes de sélection ont pour objectif la recherche d'un compromis biais/variance. L'enjeu est de trouver un sous-ensemble de covariables suffisamment grand pour que le modèle ait de bonnes qualités de prédiction et soit suffisamment petit pour éviter les redondances, le sur-ajustement et les problèmes d'inférence. Une vision différente des approches par sélection sont les méthodes de régression sur composantes et en particulier les approches de type *partial least squares* (PLS).

La technique générale de la régression PLS a été mise au point par Wold (1985) dans le but de décrire les relations entre des groupes de variables indépendantes et dépendantes dans des systèmes de type entrée-sortie comprenant de nombreuses variables. Elle a été conçue pour faire face aux problèmes résultant de l'insuffisance de l'utilisation de la régression linéaire classique, qui trouve ses limites dès lors que l'on cherche à modéliser des relations entre des variables pour lesquelles il y a peu d'individus, ou beaucoup de variables explicatives en comparaison du nombre d'individus (le nombre de variables explicatives pouvant excéder très largement le nombre d'individus), ou encore lorsque les variables explicatives sont fortement corrélées entre elles. Dans la régression PLS, le calcul des composantes f se fait en tenant compte des variables à prédire Y. Le problème revient à optimiser la covariance entre X et Y et à chercher les vecteurs u et v de norme 1 qui sont les solutions du problème d'optimisation suivant :

$$\max_{u'u=1\ ;\ v'v=1} \langle Xu|\, Yv\rangle_W$$

où W est une matrice de poids. Un des avantages de la régression PLS par rapport à des approches telles que la régression sur composantes principales est qu'elle prend en compte l'information contenue dans la réponse pour construire les composantes. En revanche il est fondamental d'utiliser un jeu de données exogène pour sélectionner le nombre d'axes pertinents pour l'analyse. De plus, les approches PLS classiques sont mal adaptées aux données qualitatives telles que les données binaires (présence/absence d'espèces) ou de comptage (abondance des espèces).

Mes contributions La « sélection de variable » en tant qu'objet de recherche n'est pas directement au centre de mes thématiques. Néanmoins, j'ai été amené à m'y intéresser de près pour comprendre en particulier le rôle de l'environnement sur l'abondance ou la distribution des espèces forestières. Cela s'est concrétisé par deux publications (Flores et al., 2009; Mortier et al., 2015) que je présenterai dans le chapitre suivant. En ce qui concerne la question de la régression sur composante, cette nouvelle thématique est le fruit d'une nouvelle collaboration avec mes collègues de l'université de Montpellier X. Bry et C. Trottier et du CIRAD G. Cornu. Cette coopération fructueuse m'a donné l'occasion de publier deux articles (Bry et al., 2013, 2015) et de développer un « package » **R** (Cornu et al., 2015).

1.5 Conclusions

Cette présentation à la fois du contexte biologique et mathématique a pour objectif d'exposer les enjeux appliqués auxquels j'ai été confronté durant ces dernières années. Cela a permis d'une part de mettre en avant quelques grands enjeux dans le contexte forestier et d'autre part de présenter les outils mathématiques auxquels j'ai fait appel pour y répondre. Je ne prétends pas avoir surmonté toutes les limites connues de ces outils, néanmoins, j'ai pris plaisir à les combiner pour en extraire les avantages et ainsi répondre aux questions multiples et variées que j'ai abordées. Le chapitre suivant présente quelques-unes de ces combinaisons pour gérer des questions de processus spatiaux multivariés non-gaussiens, la sur-représentation de zéros dans des données de comptage tout en cherchant les facteurs biotiques et abiotiques explicatifs de l'abondance d'espèces, ou encore la dynamique d'écosystèmes tropicaux riches en espèces.

2

Combiner les outils pour mieux en tirer profit

2.1 Modèles hiérarchiques bayésiens spatiaux multivariés (Chagneau et al., 2011, 2009)

2.1.1 Introduction

Plusieurs variables sont mesurées pour caractériser l'environnement : altitude, pente, drainage du sol, teneur du sol en minéraux, couleur du sol, etc. Ces variables ont été échantillonnées de manière aléatoire et en un nombre limité de sites. En chaque point d'échantillonnage, toutes les variables ont été mesurées ; ces variables ne sont pas indépendantes et il est nécessaire de se placer dans un cadre multivarié pour pouvoir prendre en compte la dépendance entre les variables. L'environnement sera donc représenté par un champ spatial multivarié. Le problème de la prédiction de champs spatiaux multivariés concerne de nombreux domaines : pédologie (McBratney et al., 2000), épidémiologie (Golam Kibria et al., 2002), économie (Gelfand et al., 2007). Ici, le champ multivarié a la particularité d'être constitué de variables de différente nature. Certaines variables comme la teneur du sol en minéraux sont continues, d'autres comme le drainage sont ordinales, d'autres encore comme la couleur du sol sont nominales. Dans certains cas, le champ spatial considéré peut également comporter des variables de comptage.

Le modèle hiérarchique spatial multivarié que nous avons développé peut être défini pour un nombre quelconque K de variables. Néanmoins, pour des raisons de simplifications, je me suis limité, ici, à un champ aléatoire composé de trois variables de nature différente : une variable gaussienne, une variable de Poisson et une variable ordinale. Le traitement de variables nominales peut également être envisagé mais complexifie inutilement la description du modèle. Le modèle est basé sur une approche hiérarchique ; tous les niveaux de la hiérarchie sont décrits successivement.

2.1.2 Modèle

Modélisation des observations, premier niveau de la hiérarchie Soient $\mathbf{x}_1, \ldots, \mathbf{x}_n$ les n sites échantillonnés. On désigne par $Y_1(\mathbf{x})$ la variable gaussienne au point $\mathbf{x}$, par $Y_2(\mathbf{x})$ la variable de Poisson au point $\mathbf{x}$ et par $Y_3(\mathbf{x})$ la variable ordinale à L modalités au point $\mathbf{x}$. Soit $\mathbf{Y}_k = (Y_k(\mathbf{x}_1), \ldots, Y_k(\mathbf{x}_n))'$, $k = 1, 2, 3$ le vecteur de la variable Y_k en chacun des sites échantillonnés. Soit $\mathbf{Y} = (\mathbf{Y}_1{}', \mathbf{Y}_2{}', \mathbf{Y}_3{}')'$ le vecteur de toutes les variables mesurées en tous les sites.

La variable gaussienne $Y_1(\mathbf{x})$ et la variable de Poisson $Y_2(\mathbf{x})$ dépendent de variables latentes $E_1(\mathbf{x})$ et $E_2(\mathbf{x})$. Les variables $E_1(\mathbf{x})$ et $E_2(\mathbf{x})$ sont les composantes spatiales intervenant dans le modèle linéaire généralisé associé à chaque variable $Y_1(\mathbf{x})$ et $Y_2(\mathbf{x})$. Conditionnellement à $E_1(\mathbf{x})$ et $E_2(\mathbf{x})$, les variables $Y_1(\mathbf{x})$ et $Y_2(\mathbf{x})$ sont indépendantes. Pour les variables gaussiennes et de Poisson, nous suivons donc le modèle linéaire généralisé proposé par Diggle et al. (1998) :

$$Y_1(\mathbf{x}_i)|\mu_1, E_1(\mathbf{x}_i), \nu_1 \sim \mathcal{N}(\mu_1 + E_1(\mathbf{x}_i), \nu_1^2), \tag{2.1}$$

$$Y_2(\mathbf{x}_i)|\mu_2, E_2(\mathbf{x}_i) \sim \mathcal{P}\{\exp(\mu_2 + E_2(\mathbf{x}_i))\} \tag{2.2}$$

où $\mathcal{P}(\lambda)$ désigne la loi de Poisson de paramètre λ. Les paramètres μ_1 et μ_2 représentent les effets moyens des variables Y_1 et Y_2. Le paramètre ν_1^2 correspond à l'effet de pépite associé à la variable gaussienne Y_1.

La modélisation de la variable ordinale se fonde sur les modèles probit, introduits dans le cas univarié par Bliss (1935), appartiennent à la classe des modèles linéaires généralisés (McCullagh and Nelder, 1989). Ils permettent de modéliser la relation entre la probabilité de réponse d'une variable discrète et un ensemble de variables explicatives. Les modèles probit peuvent être définis en terme de variables sous-jacentes gaussiennes Z (Albert and Chib, 1993; Chib and Greenberg, 1998; Chen and Shao, 1999; Bar-Hen and Mortier,

2004; Chaubert et al., 2008). La généralisation au cas spatial est directe en modélisant la dépendance entre les variables sous-jacentes Z :

$$\mathbb{P}(Y_3(\mathbf{x}_i) = j|\boldsymbol{\alpha}_3, E_3(\mathbf{x}_i), \mu_3) = \mathbb{P}(Z_3(\mathbf{x}_i) \in]\alpha_{3;\,j-1}, \alpha_{3;\,j}]|E_3(\mathbf{x}_i), \mu_3), \quad (2.3)$$

$$Z_3(\mathbf{x}_i)|E_3(\mathbf{x}_i), \mu_3 \sim \mathcal{N}(\mu_3 + E_3(\mathbf{x}_i), 1). \quad (2.4)$$

$\boldsymbol{\alpha}_3 = (-\infty, 0, \alpha_{3,2}, \ldots, \alpha_{3,L-1}, +\infty)$ désigne le vecteur des seuils relatifs à la variable gaussienne sous-jacente $Z_3(\mathbf{x})$. Le paramètre μ_3 est l'effet moyen associé à la variable $Z_3(\mathbf{x})$. Conditionnellement à $E_1(\mathbf{x}_i)$ (respectivement $E_2(\mathbf{x}_i)$) et $E_3(\mathbf{x}_j)$, les variables $Y_1(\mathbf{x}_i)$ (respectivement $Y_2(\mathbf{x}_i)$) et $Y_3(\mathbf{x}_j)$ sont indépendantes. Les expressions (2.1) à (2.4) constituent le premier niveau du modèle hiérarchique.

Modèles spatiaux mulivariés pour processus non gaussiens, deuxième niveau de la hiérarchie Le second niveau du modèle hiérarchique permet de décrire la structure de dépendance entre les variables. La dépendance spatiale entre les processus $Y_k(.)$ est portée par les variables latentes $E_k(\mathbf{x})$, $k = 1, 2, 3$. Les variables $E_k(\mathbf{x})$ sont construites suivant la méthode moyenne mobile proposée par Ver Hoef and Barry (1998), c'est-à-dire par convolution d'une fonction dite moyenne mobile avec un mélange de bruits blancs.

Soit V_k, $k = 1, 2, 3$ une combinaison linéaire de bruits blancs

$$V_k(\mathbf{x}|\rho_k, \boldsymbol{\Delta}_k) = \sqrt{1 - \rho_k^2} W_k(\mathbf{x}) + \rho_k W_0(\mathbf{x} - \boldsymbol{\Delta}_k)$$

où $W_k(.)$, $k = 1, 2, 3$ est un bruit blanc (Yaglom, 1987), ρ_k appartient à l'intervalle $[-1; 1]$ et où $\boldsymbol{\Delta}_k$ appartient à $\mathbb{R}^2$. Le processus $W_0(.)$ induit une dépendance entre les processus $V_k(.)$ puisque, pour tout $k \neq m$

$$\text{Cor}\left[\int_{\mathbb{R}^2} V_k(\mathbf{x} + \boldsymbol{\Delta}_k|\rho_k, \boldsymbol{\Delta}_k)\mathrm{d}\mathbf{x}, \int_{\mathbb{R}^2} V_m(\mathbf{x} + \boldsymbol{\Delta}_m|\rho_m, \boldsymbol{\Delta}_m)\mathrm{d}\mathbf{x}\right] = \rho_k\rho_m \equiv \rho_{km}.$$

Le paramètre ρ_{km} peut être considéré comme la corrélation croisée entre les mélanges de bruits blancs V_k et V_m (Yaglom, 1987; Ver Hoef and Barry, 1998). Soit f_k, $k = 1, 2, 3$ une fonction moyenne mobile définie sur $\mathbb{R}^2$. Soit $\boldsymbol{\theta}_k$ le vecteur des paramètres associés à f_k. La variable aléatoire $E_k(\mathbf{x}_i)$ est définie par

$$E_k(\mathbf{x}_i) = \int_{\mathbb{R}^2} f_k(\mathbf{u} - \mathbf{x}_i|\boldsymbol{\theta}_k) V_k(\mathbf{u}|\rho_k, \boldsymbol{\Delta}_k)\mathrm{d}\mathbf{u}.$$

Les variables $E_k(\mathbf{x}_i)$, $i = 1, \ldots, n$ sont dépendantes car les processus $V_k(.)$ le sont. La distribution conditionnelle $\mathbf{E} = (\mathbf{E}_1{}', \mathbf{E}_2{}', \mathbf{E}_3{}')'$ est une loi normale multivariée de moyenne nulle et de matrice de covariance $\mathbf{C}$:

$$\mathbf{E}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}, \boldsymbol{\Delta} \sim \mathcal{N}_{3n}(\mathbf{0}, \mathbf{C})$$

où $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$ et $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_3)$. Cela constitue le second niveau de la hiérarchie. L'un des avantages de la construction moyenne mobile est que l'expression de la matrice de covariance $\mathbf{C}$ est connue :

$$C_{kk}(\mathbf{h}) = \mathrm{Cov}[E_k(\mathbf{x}), E_k(\mathbf{x}+\mathbf{h})] = \int_{\mathbb{R}^2} f_k(\mathbf{u}) f_k(\mathbf{u}-\mathbf{h}) \mathrm{d}\mathbf{u}, \tag{2.5}$$

$$C_{km}(\mathbf{h}) = \mathrm{Cov}[E_k(\mathbf{x}), E_m(\mathbf{x}+\mathbf{h})] = \rho_k \rho_m \int_{\mathbb{R}^2} f_k(\mathbf{u}) f_m(\mathbf{u}-\mathbf{h}+\boldsymbol{\Delta}_m-\boldsymbol{\Delta}_k) \mathrm{d}\mathbf{u}. \tag{2.6}$$

Selon les fonctions moyennes mobiles choisies, le calcul des intégrales peut être explicite ou non.

Inférence L'algorithme que nous avons développé pour estimer les paramètres du modèle est basé sur un algorithme de Monte Carlo par chaînes de Markov (MCMC). Celui-ci combine dans une même procédure des échantillonneurs de Gibbs, de Métropolis-Hastings ou encore des méthodes adaptatives, tel que l'algorithme de Langevin-Hastings tronqué. Appelé aussi en anglais « Metropolis-adjusted Langevin algorithm » (MALA), il a été introduit par Besag (1994); Atchade and Rosenthal (2005); Atchade (2006), puis étudié plus en détail par Roberts and Tweedie (1996). C'est un algorithme de Metropolis-Hastings pour lequel la loi de proposition est donnée par :

$$\mathcal{N}_d\left(\mathbf{x} + \frac{\sigma^2}{2}\nabla \ln \pi(\mathbf{x}), \sigma^2 \mathrm{I}_d\right)$$

où d est la dimension de l'espace des paramètres et où le terme $\nabla \ln \pi(\mathbf{x})$, appelé dérive, désigne le gradient de $\ln \pi(\mathbf{x})$. L'utilisation du gradient dans la loi de proposition permet d'obtenir de meilleures propriétés de convergence qu'avec un algorithme de Metropolis-Hastings où la loi de distribution serait $\mathcal{N}_d(\mathbf{x}, \sigma^2 \mathrm{I}_d)$ (Christensen et al., 2001; Christensen and Waagepetersen, 2002). La variance de proposition σ^2 ($\sigma > 0$) est spécifiée par l'utilisateur (Møller and Waagepetersen, 2004). Des résultats théoriques obtenus par Roberts and Rosenthal (1998) et Breyer and Roberts (2000) suggèrent de choisir σ de façon à obtenir un taux d'acceptation qui soit à peu près égal à 0,574. Pour éviter des problèmes de dégénérescence dans le taux de convergence de l'algorithme, on se limite, en général, à l'utilisation de dérives qui soient des fonctions bornées. En pratique, la façon la plus simple d'obtenir une fonction de dérive bornée est de tronquer la fonction prise pour dérive. Nous obtenons alors une version tronquée de l'algorithme MALA en remplaçant la quantité $\nabla \ln \pi(\mathbf{x})$ dans la loi de proposition par :

$$D_{MALA}(\mathbf{x}) = \frac{\delta}{\max(\delta, |\nabla \ln \pi(\mathbf{x})|)} \nabla \ln \pi(\mathbf{x})$$

où $\delta > 0$ est une constante fixée.

2.1.3 Conclusions et perspectives

L'objectif de ce travail était de proposer un modèle spatial multivarié original qui permette de prédire des variables de différente nature. Il a été réalisé au cours de la thèse de Pierrette Chagneau qui a été ma première doctorante. Ce modèle spatial multivarié hiérarchique permet de traiter simultanément des variables gaussiennes, de Poisson et ordinales ou multinomiales, grâce à une approche basée sur les modèles linéaires généralisés spatiaux. Les simulations réalisées, que je n'ai pas présentées ici, ont confirmé la capacité du modèle à prédire différents types de variables et ont montré qu'une procédure d'estimation multivariée conduit à des prédictions de meilleure qualité qu'une procédure d'estimation univariée. Dans le modèle, la dépendance entre les variables se traduit au travers de la dépendance de leurs composantes spatiales S_k. La structure de dépendance est modélisée par une matrice obtenue grâce à la construction moyenne mobile. Remarquons qu'il est possible d'utiliser un modèle de covariance classique, mais, en l'absence d'information sur les variables latentes S_k, l'approche moyenne mobile offre l'avantage d'être plus flexible.

Néanmoins, le choix des fonctions moyennes mobiles est délicat et reste un problème ouvert. Il pourrait être intéressant de tester la robustesse des prédictions selon le type et la forme des fonctions moyennes mobiles grâce à des simulations. Une extension de la construction moyenne mobile pourrait être envisagée afin de prendre en compte la dépendance spatiale à différentes échelles comme dans le modèle linéaire de corégionalisation. De plus, la procédure d'estimation des paramètres est gourmande en ressources informatiques et le temps de calcul nécessaire à l'estimation peut s'avérer long. Ce dernier augmente avec le nombre de variables étudiées et le nombre de sites échantillonnés, car les matrices manipulées sont alors de grande dimension. Plusieurs alternatives peuvent être envisagées pour remédier à ce problème. Il pourrait être fait appel à des méthodes basées sur la vraisemblance composite (Varin, 2008). Une autre alternative consisterait à simplifier la procédure d'estimation en suivant l'approche proposée par Joe (1997). Cette approche consiste, dans un premier temps, à effectuer autant de procédures d'estimation univariée qu'il y a de variables dans le modèle, afin d'estimer les paramètres relatifs à chacune des variables. Dans un second temps, une procédure d'estimation multivariée est lancée. Les paramètres associés à chacune des variables sont considérés comme connus, seul le vecteur de corrélations $\boldsymbol{\rho}$ est estimé, ce qui permet de réduire la durée des calculs. Enfin, une autre solution consisterait à utiliser une inférence bayésienne approchée (Rue et al.,

2009) au lieu des simulations MCMC.

2.2 Modèles hierarchiques bayésiens spatiaux avec sur-représentation de zéros et sélection de variables (Flores et al., 2009)

2.2.1 Introduction

La régénération, qui désigne l'ensemble des processus allant de la floraison d'un arbre adulte à l'apparition d'un nouvel individu dans le peuplement, est une composante fondamentale pour comprendre la dynamique forestière. La répartition spatiale des juvéniles dépend des mécanismes de dispersion des graines et de la position géographique des adultes mais aussi de l'environnement. En fonction de leur niche écologique et de leur plasticité, les essences sont susceptibles ou non de s'installer dans différents types d'environnement. Plusieurs modèles mathématiques ont été développés pour prédire la régénération. Certains modèles ne permettent de déterminer que le nombre d'arbres recrutés (Vanclay, 1992; Lexerød, 2005), d'autres plus complexes permettent de mieux comprendre les mécanismes de dispersion ou la survie des juvéniles (Sagnard et al., 2007; Eerikäinen et al., 2007). L'information génétique aussi a été mise à profit pour améliorer la compréhension des mécanismes de dispersion (Jones and Muller-Landau, 2008). Enfin, certains modèles se focalisent sur la distribution spatiale des juvéniles. C'est l'exemple que je présente maintenant. Le travail de thèse d'O. Flores avait entre autres pour objectif de modéliser l'abondance des juvéniles de 6 espèces tropicales de Guyane française en fonction de facteurs environnementaux. Nous avons développé un modèle qui permet de prendre en compte simultanément (i) la sur-représentation de zéros, (ii) la dépendance spatiale entre les quadrats et (iii) la sélection des variables environnementales pertinentes. Nous nous sommes placés à nouveau dans un cadre hiérarchique bayésien.

2.2.2 Modèle

Les données d'abondance sont classiquement modélisées par des lois de Poisson. Mais cette approche suppose que l'espérance est égale à la variance. Or en écologie comme dans d'autres domaines d'application, cette hypothèse n'est pas valide en règle générale. On observe fréquemment une sur-dispersion (plus rarement une sous-dispersion). Une approche classique consiste à considérer une loi négative binomiale. Mais une cause particulière

de dispersion statistique est l'excès de zéros, c'est-à-dire de données de comptage nul, par rapport à une distribution de Poisson (ou zéro-inflation). Une méthode consiste à supposer que la loi des données est un mélange de deux lois de Poisson simples, l'une d'espérance nulle $\mathcal{P}(0)$ (masse de dirac en zéro) et l'autre d'espérance strictement positive $\mathcal{P}(\mu)$. La proportion de mélange entre les deux lois est déterminée par une loi de Bernoulli $\mathcal{B}$ de paramètre ω inconnu.

On dit que Z suit une loi *zero-inflated Poisson* (ZIP), si

$$\begin{aligned} \mathbb{P}(Z = z|\omega, \mu) &= \begin{cases} \omega + (1-\omega)\mathbb{P}(Z = 0|\mu), & \text{if } z = 0 \\ (1-\omega)\mathbb{P}(Z \neq 0|\mu), & \text{if } z > 0 \end{cases} \\ &\Updownarrow \\ f(z|\omega, \mu) &= \omega\mathcal{P}(0) + (1-\omega)\mathcal{P}(\mu) \end{aligned}$$

De manière hiérarchique un modèle ZIP se définit par l'introduction d'un vecteur latent C. Soit $C = (C_1, \ldots, C_n)$ un vecteur aléatoire latent tel que C_i soit égal à $c_i = 0$ si $Z_i > 0$ ou si Z_i est nul et issu de $\mathcal{P}(\mu)$ et $c_i = 1$ si $Z_i = 0$ et issu de $\mathcal{P}(0)$. La distribution des éléments C_i de C est une loi de Bernoulli de paramètre ω. La loi jointe de Z et C est :

$$\begin{aligned} \ell(Z, C|\omega, \mu) &= \prod_{i=1}^{n} \ell(Z_i|C_i = c_i, p, \mu)\pi(C_i|p) \\ &= \prod_{i=1}^{n} \omega^{c_i} \left[(1-\omega)\, P(\mu)\right]^{1-c_i} \end{aligned}$$

Le deuxième niveau de la hiérarchie permet de modéliser les paramètres de la loi de Poisson et de la loi de Bernoulli en fonction de variables environnementales, comme on le fait dans des modèles linéaires généralisés. La différence est que nous proposons de modéliser l'intensité de la loi de Poisson conditionelement à un processus spatial $\alpha(s)$. A la différence du modèle spatial présenté dans le chapite pécédent, nous souhaitons considérer ici un processus spatial sur lattice. Le modèle proposé est un modèle *conditional autoregressive* (CAR) (Banerjee et al., 2004).

$$\alpha(s_i)|\alpha(s_j) \underset{i\sim j}{\sim} \mathcal{N}\left(\rho \sum_{j\sim i} b_{ij}\alpha(s_j), \sigma_i\right) \tag{2.7}$$

où ρ décrit la force de dépendance spatiale, entre les voisins et σ un paramètre d'échelle. Le processus spatial est déterminé conditionnellement à un voisinage donné. $j \sim i$ décrit la relation de voisinage des sites i et j et b_{ij} un système de pondération connu.

Donc finalement, on modélise l'intensité de la loi de Poisson et la probabilité de la loi de Bernoulli comme

$$\begin{aligned} \text{logit}[\omega(\mathbf{s})|\xi] &= \mathbf{B}\xi \\ \log[\lambda(\mathbf{s})|\beta, \alpha(\mathbf{s})] &= \mathbf{P}\beta + \alpha(\mathbf{s}) \end{aligned} \tag{2.8}$$

où $\mathbf{B}$ et $\mathbf{X}$ sont des matrices de design connues, ξ et β des paramètres inconnus et où $\text{logit}(x) = \log(x/(1-x))$.

Dans ce travail nous souhaitions aussi sélectionner les variable environnementales pertinentes. Pour ce faire, nous avons introduit deux vecteurs aléatoires supplémentaires, γ et η composés chacun de 1 ou 0. Si $\gamma_i = 1$ cela traduit le fait que la variable P_i est inclus dans le modèle et si $\gamma_i = 0$ le rôle de cette variable, sur l'intensité de la loi de Poisson, est nul. De manière similaire si $\eta_j = 1$, la co-variable B_j est inclue dans le modèle associé à ω, si $\eta_j = 0$, elle ne l'est pas. Cette approche est décrite précisément par exemple dans Ntzoufras et al. (2000); Dellaportas et al. (2002). Ainsi les équations 2.8 se réécrivent de la façon suivante pour une observation $i = 1, \ldots, n$

$$\begin{aligned} logit[\omega(\mathbf{s_i})|\xi, \eta] &= \sum_{j}^{p} \mathbf{B}_{ij}\eta_j\xi_j \\ log[\lambda(\mathbf{s_i})|\beta, \gamma, \alpha(\mathbf{s})] &= \sum_{j} \mathbf{P}_{ij}\gamma_j\beta_j + \alpha(\mathbf{s_i}) \end{aligned} \tag{2.9}$$

2.2.3 Conclusions et perspectives

Ce travail a été réalisé au cours du doctorat d'Olivier Flores. Nous avons aussi construit un algorithme MCMC adaptatif pour estimer les paramètres du modèle. Bien qu'il soit possible d'utiliser le logiciel libre *openBugs*, nous avons programmé dans son ensemble le modèle ci-dessus. Cela a permis de gagner grandement en temps de calcul et de tester plus facilement l'influence des choix des lois *a priori*. L'avantage de cette approche réside dans sa généralité mais impose une contrainte forte qui fait encore débat dans la communauté : comment modéliser de façon spatialement explicite chaque terme du modèle (présence/abondance) et traiter simultanément l'ensemble des espèces du peuplement forestier ?

2.3 Modèles de mélange pour grouper les espèces selon leurs dynamiques (Ouédraogo et al., 2013; Mortier et al., 2013, 2015)

2.3.1 Introduction

Comprendre la dynamique d'un écosystème forestier naturel nécessite de modéliser la dynamique de l'ensemble des espèces ou genres. Les modélisateurs forestiers ont développé un grand nombre de modèles pour comprendre l'évolution des arbres et des peuplements. Les modèles de dynamique forestière peuvent être classés en trois groupes suivant le niveau de complexité du peuplement sur lequel ils reposent. On distingue (Vanclay, 1995) :

- *les modèles globaux :* ne mettent en jeu que des variables qui décrivent globalement la population (densité, diamètre moyen). Chaque arbre est considéré comme une réalisation de l'arbre moyen du peuplement. Ces modèles ne prennent pas en compte l'hétérogénéité entre individus au sein de la population. Ils sont particulièrement adaptés pour décrire des peuplements dits « homogènes" ou « réguliers" (c'est-à-dire monospécifiques et équiennes).
- *les modèles individuels : décrivent le peuplement sur la base du niveau individu.* La trajectoire de la variable étudiée est suivie pour chaque arbre. Les modèles individuels sont spatialisés ou non. Ce type de modèles permet de prendre en compte l'hétérogénéité au sein du peuplement.

  Les modèles individuels sont basés sur trois grandes composantes synthétisant la démographie du peuplement :
  - un modèle de croissance intégrant l'effet du milieu et de la compétition (interactions entre arbres),
  - un modèle de mortalité décrivant pour chaque individu la probabilité de mourir ou de survivre en fonction de différents facteurs biotiques ou abiotiques,
  - un modèle de régénération ou de recrutement décrivant l'apparition de nouveaux individus dans le peuplement.
- *les modèles de distribution* Ces modèles forment un ensemble d'outils qui présentent l'avantage d'être à la fois plus simples que les modèles individus-centrés mais aussi plus souples que les modèles globaux. Dans ces modèles, la population n'est plus décrite par une variable moyenne comme dans les modèles globaux, mais est résumée par une fonction de distribution sur une ou plusieurs variables (diamètre, hauteur de

l'arbre, etc). La modélisation consiste à suivre l'évolution de cette fonction dans le temps (le temps sera ici discret). Les modèles matriciels (Caswell, 2001) qui correspondent à des modèles à espace d'états et temps discret sont les plus utilisés en foresterie. En particulier les modèles de Usher sont adaptés à des populations structurées en taille (Usher, 1966, 1969). Ces modèles décrivent la dynamique d'un vecteur d'effectif $\mathbf{N}(t)$ dont chaque élément $N_{l,t}$ représente le nombre d'individus dans L classes de diamètre ordonnées $l = 1, \ldots, L$. Le modèle de Usher initial se base sur quatre hypothèses

1. l'hypothèse d'indépendance : les individus ont des dynamiques indépendantes (la dynamique d'un individu n'est pas influencée par celles des autres individus, ce qui signifie qu'on ne prend pas en compte la compétition par exemple).
2. l'hypothèse de Markov : les effectifs au temps $t+1$ ne dépendent que des effectifs au temps t (la croissance d'un arbre et donc sa probabilité de changer de classe de taille ne dépend pas de son diamètre initial ni de sa croissance passée).
3. l'hypothèse de Usher : entre deux temps t et $t+1$, un arbre ne peut ni passer dans une classe de diamètre inférieure, ni franchir plus d'une classe de diamètre.
4. l'hypothèse de stationnarité : les probabilités de transition sont constantes au cours du temps.

Tout comme chaque modèle matriciel, le modèle de Usher peut s'interpréter en terme d'espérance d'une chaîne de Markov homogène :

$$\mathbb{E}[\mathbf{N}_{t+1}|\mathbf{N}_t] = U\,\mathbf{N}_t \tag{2.10}$$

où U, la matrice de transition appelée matrice de Usher est égale à

$$U = \begin{pmatrix} p_1 + f & f & \cdots & f \\ q_1 & p_2 & & 0 \\ & \ddots & \ddots & \\ 0 & & q_{L-1} & p_L \end{pmatrix} \tag{2.11}$$

p_l est la probabilité d'un individu de rester dans le classe l, q_l de passer d'une classe l à $l+1$ entre t et $t+1$ et f le nombre de nouveaux recrutés. q_l et p_l appartiennent à $[0,1]$, tandis que f appartient à $\mathbf{R}^+$. La probabilité de mourir d'un individu dans la classe l est égal à $m_l = 1 - p_l - q_l$. Posons $\mathbf{d} = (d_1, \ldots, d_L)$ la distribution des classes diamétriques dans la population de sorte que d_l soit la probabilité de tirer un individu

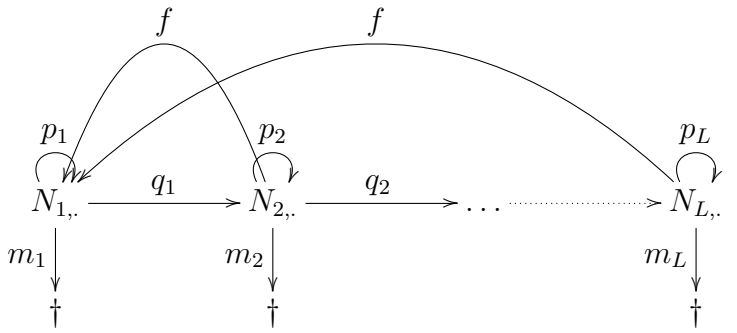


FIGURE 2.1: Représentation du cycle de vie modélisé par un modèle de Usher. p_l est la probabilité pour un individu de rester dans la classe l, q_l celle de passer de la classe l à $l+1$, m_l la probabilité de mourir et enfin f est le taux de fécondité.

dans la population appartenant à la classe l ($\sum_{l=1}^{L} d_l = 1$), $N_{l,l,t}$ le nombre d'individus de la classe l et qui y reste entre $t-1$ et t, $N_{l,l+1,t}$ le nombre d'individus qui passe de la classe l à la classe $l+1$ entre $t-1$ et t, et $N_{l,\dagger,t}$ le nombre d'individus qui meurent dans la classe l entre $t-1$ et t. Enfin, soit R_t le nombre d'événements de recrutement entre $t-1$ et t, supposé distribué selon une loi de Poisson d'intensité fN_{t-1}. Le vecteurs $\mathbf{N} = (N_{1,l,t}, \ldots, N_{L,\dagger,t}, \mathbf{N}_{t-1}, R_t)$ constitue les observations. Graphiquement un modèle de Usher peut se représenter ainsi 2.1

2.3.2 Modèles de Usher homogène en mélange (Mortier et al., 2013)

Introduction

Dans cet exemple, je présente comment nous utilisons le cadre des modèles de mélange combiné aux modèles de Usher pour proposer une méthode de prédiction de la dynamique d'un écosystème composé d'un grand nombre d'espèces. En particulier nous proposons une méthode de classification en groupes d'espèces et utilisons ces groupes pour prédire la dynamique de l'écosystème dans son ensemble. Nous nous sommes placés dans ce travail sous un angle bayésien.

Modèle

D'un point de vue statistique, pour une population, la vraisemblance jointe associée au modèle de Usher et ainsi à la chaîne de Markov sous-jacente

est donnée par

$$\begin{aligned}\mathcal{L}(N|\theta) &= \prod_{l=1}^{L-1} \mathcal{M}(N_{l,l,t}, N_{l,l+1,t}, N_{l,\dagger,t}|p_l, q_l, m_l, N_{l,t-1}) \\ &\quad \times \mathcal{M}(N_{L,L,t}, N_{L,\dagger,t}|p_L, m_L, N_{L,t-1}) \\ &\quad \times \mathcal{M}(N_{1,t-1}, \ldots, N_{L,t-1}|d_1, \ldots, d_L, N_{t-1}) \\ &\quad \times \mathcal{P}(R_t|fN_{t-1})\end{aligned} \tag{2.12}$$

où $\mathcal{M}$ représente la loi Multinomiale, $\mathcal{P}$ la loi de Poisson, et $\theta = (p,\, q,\, m,\, f,\, d)$ le vecteur des paramètres où $p = (p_1,\, \ldots,\, p_L)$, $q = (q_1,\, \ldots,\, q_{L-1})$ et $m = (m_1,\, \ldots,\, m_L)$.

Supposons maintenant que la population soit issue de K groupes d'espèces de sorte que chaque sous-population soit régie par sa propre dynamique modélisée par un modèle de Usher spécifique. Ainsi, on peut supposer qu'il existe K matrices de Usher $U_1,\, \ldots,\, U_K$. Comme l'assignation de chaque espèce à un groupe n'est pas connue *a priori*, on définit une variable aléatoire latente C qui identifie l'appartenance de chaque espèce à son groupe. Par exemple si l'espèce s appartient au troisième groupe, $C_s = 3$. La prédiction de la dynamique de l'espèce s sera alors construite en remplaçant la matrice de Usher U donnée par l'equation 2.10 par la matrice correspondant au groupe $3 : U_3$. Néanmoins pour tenir compte des incertitudes liées à la classification et à son estimation, la prédiction de la population peut se réécrire de la façon suivante :

$$\mathbb{E}[N_{t+1}|N_t] = \sum_{k=1}^{K} \pi_k\, U_k N_t \tag{2.13}$$

où $\pi_k$est la probabilité *a priori* que C soit égal à k. L'équation 2.13 définit ce que nous avons appelé le mélange de modèles matriciels de Usher homogène et dont la vraisemblance est égale à :

$$\mathcal{L}(N|\theta, \pi) = \sum_{k=1}^{K} \pi_k\, \mathcal{L}(N|\theta_k) \tag{2.14}$$

où $\theta = (\theta_1,\, \ldots,\, \theta_K)$ est le vecteur des paramètres, θ_k le vecteur associé au $K^{\text{ème}}$ modèle matriciel, $\pi = (\pi_1,\, \ldots,\, \pi_K)$ le vecteur des probabilités *priori*, et $\mathcal{L}(N|\theta_k)$ est donné par l'équation 2.12. Les espèces peuvent être assignées à un groupe g en utilisant le maximum *a posteriori* : $\pi_g = \max_k\{\pi_k\}$.

Lois *a posteriori* et *a priori*

Soit S le nombre d'espèce. On pose $N^s = (N^s_{1,l,t}, \ldots, N^s_{L,\dagger,t}, N^s_{t-1}, R^s_t)$ le vecteur des observations pour l'espèce $s = 1, \ldots, S$ et $\mathbf{N} = (N^1, \ldots, N^S)$ le

vecteur des observations pour toutes les espèces. Soit $C = (C_1, \ldots, C_S)$ le vecteur de classification latent qui assigne chaque espèce à un groupe. En considérant K inconnu, la loi *a posteriori* s'écrit :

$$\pi^{\mathbf{N}}_{C,\theta,K}(C, \theta, K|\mathbf{N}) \propto \prod_{s=1}^{S} \mathcal{L}(N^s|\theta_{C_s})\ \pi^0_{C|\theta,K}(C|\theta, K)\ \pi^0_{\theta|K}(\theta|K)\ \pi^0_K(K) \quad (2.15)$$

où $\mathcal{L}(N^s|\theta_{C_s})$ est donné par l'équation 2.12, et $\pi^0_{C|\theta,K}$, $\pi^0_{\theta|K}$ et π^0_K désignent les lois *a priori* associées au vecteur latent C, aux paramètres des modèles de Usher et au nombre de groupes. Nous proposons dans ce travail les lois *a priori* suivantes :

- On suppose que π^0_K est une loi de Poisson de moyenne 1 strictement positive (non nulle) : $\pi^0_K(K) \equiv \mathcal{P}(1)\backslash\{0\}$. Cette loi *a priori* a été proposée par Richardson and Green (1997); Nobile (2005). Celle-ci permet de « limiter » le nombre de groupes estimés et ainsi d'être plus parcimonieux comparé à la loi uniforme $\mathcal{U}[1, \ldots, S]$ classiquement utilisée.
- On suppose que $\pi^0_{\theta|K}$, loi *a priori* des paramètres des différentes classes et des différents groupes sont indépendantes :

  $$\pi^0_{\theta|K}(\theta|K) = \prod_{k=1}^{K} \left\{ \prod_{l=1}^{L-1} \pi^0_{p,q,m|l,k}(p_{lk}, q_{lk}, m_{lk}) \right\} \pi^0_{p,m|k}(p_{Lk}, m_{Lk})\ \pi^0_{\mathbf{d}|k}(\mathbf{d}_k)\ \pi^0_{f|k}(f_k)$$

  On utilise ici des lois conjuguées, Dirichlet pour les lois multinomiales et Gamma pour les lois de Poisson, de sorte que
  - $\pi^0_{\mathbf{d}|k} \equiv \mathcal{D}(\alpha, \ldots, \alpha)$, où les hyper-paramètres α sont fixés à 1 (loi uniforme)
  - $\pi^0_{p,q,m|l,k} \equiv \mathcal{D}(\beta, \beta, \beta)$ et $\pi^0_{p,m|k} \equiv \mathcal{D}(\beta, \beta)$, où les hyper-paramètres β sont fixés à 1 (loi uniforme)
  - $\pi^0_{f|k} \equiv \mathcal{G}(\gamma, \delta)$, où δ et γ sont les hyper-paramètres égaux à $\gamma = 0.01$ et $\delta = 1$ pour tenir compte de la connaissance des experts sur le taux moyen de reproduction d'une espèce qui est approximativement égal à 1% de ses effectifs .
- on suppose que chaque espèce, indépendamment des autres espèces, a une probabilité uniforme d'appartenir à un groupe. Ainsi la loi *a priori* du vecteur latent C est donnée par

  $$\pi^0_{C|\theta,K}(C|\theta, K) = \prod_{s=1}^{S} \pi^0_{C|K}(C_s|K)$$

  où $\pi^0_{C|K}(C_s|K)$ est la loi uniforme sur le nombre de groupe : $\mathcal{U}(1, \ldots, K)$.

Inférence et algorithme

L'inférence des paramètres se fait par l'étude de la loi *a posteriori* $\pi^{\mathbf{N}}_{C,\theta,K}(C, \theta, K|\mathbf{N})$ définie par l'équation 2.15. Mais comme le nombre de groupes K et l'assignation des espèces aux différents groupes n'est pas connue, il n'existe pas de forme analytique simple et d'algorithme simple. Nous proposons un algorithme de type *Reversible Jump MCMC* (RJ-MCMC) (Richardson and Green, 1997). Cet algorithme consiste à réaliser trois mouvements : (i) augmenter de un le nombre de groupes, (ii) diminuer de un le nombre de groupes et (iii) garder le nombre de groupes constant mais en changeant le groupe d'une espèce. Chaque mouvement au cours de l'algorithme est choisi aléatoirement avec une probabilité 1/3. L'algorithme est le suivant

1. Proposition. Soit $|k|$ le nombre d'espèces dans le groupe $k, k = 1, \ldots, K$. Soit $K^\star$ le nouveau nombre de groupes ($k^\star = K+1$ ou $K-1$ ou encore K) et on note $C^\star$ le vecteur d'assignation associé.
   - Cas sans changement de dimension : $K^\star = K$. On propose $\mathbf{C}^\star = (C^\star_1, \ldots, C^\star_S)$ selon les deux étapes suivantes :
     (a) on choisit aléatoirement une espèce s dans un groupe qui contient au moins deux espèces ;
     (b) on propose un nouvel assignement de l'espèce s : $C^\star_s$ selon la loi multinomiale $\mathcal{M}(1\,;\, w_1, \ldots, w_K)$, tandis que $C^\star_t = C_t$ pour tous les autres espèces $t \neq s$. Les coefficients w_k sont égaux à
     $$w_k = \frac{\mathcal{L}(\mathbf{N}^s|\theta_k)}{\sum_{j=1}^K \mathcal{L}(\mathbf{N}^s|\theta_j)}$$
     où $\mathcal{L}$ est donnée par l'équation (2.12).
   - Naissance : $K^\star = K+1$. Le nouveau vecteur d'assignation $C^\star$ est obtenu en découpant un groupe choisi aléatoirement parmi les groupes qui contiennent au moins deux espèces :
     (a) choisir un groupe k parmi les groupes ayant au moins deux espèces. Les deux sous-groupes sont labellisés k_1 et k_2 ;
     (b) choisir aléatoirement $|k_1|$ le nombre d'espèces du groupe k qui composeront le groupe k_1 selon la loi uniforme $|k_1| \sim \mathcal{U}(1, \ldots, |k|-1)$
     (c) choisir $|k_1|$ espèces parmi les k. Les autres formeront le groupe k_2. Soit D le nouveau vecteur d'allocation des $|k|$ espèces réparties entre k_1 et k_2.

$\mathbf{C}^\star$ est alors égal à $(C, k, |k_1|, D)$. La loi conditionnelle de la nouvelle classification dans $K+1$ groupes sachant C et K, $\pi^{\text{split}}_{C^\star|C,K}$, est définie par :

$$\begin{aligned}\pi^{\text{split}}_{C^\star|C,K}(C^\star|C,K) &= \Pr(C^\star = (C,k,|k_1|,D)|C,K) \\ &= \frac{|k_1|!(|k|-|k_1|)!}{|k|!}\frac{1}{|k|-1}\frac{1}{\sum_{i=1}^{K}\mathbb{1}_{|i|>1}}\frac{1}{3}\end{aligned}$$

– Mort : $K^\star = K-1$. Pour proposer un nouveau vecteur d'assignation on propose simplement de fusionner deux groupes tirés aléatoirement. Donc si k_1 et k_2 sont les deux groupes sélectionnés, on note $\mathbf{C}^\star = (\mathbf{C}, k_1, k_2)$ le nouveau vecteur d'assignation. Ainsi la probabilité de cette nouvelle classification sachant l'état initial est

$$\begin{aligned}\pi^{\text{merge}}_{C^\star|C,K}(C^\star|C,K) &= \Pr(C^\star = (C,k_1,k_2)|C,K) \\ &= \frac{2!(K-2)!}{K!}\frac{1}{3}\end{aligned}$$

2. Mise à jour des paramètres connaissant C et K : les nouveaux paramètres $\theta^\star = (p^\star, q^\star, m^\star, f^\star, d^\star)$ sont échantillonnés dans les lois marginales *a posteriori* (Marin and Robert, 2007).

Les équations suivantes présentent les ratios d'acceptation/rejection de l'algorithme de Metropolis-Hastings dans le cas de la diminution du nombre de groupes (mort) : $K^\star = K-1$. Supposons que les 2 groupes k_1 et k_2 aient été choisis et concaténés dans le groupe k. Alors,

$$\frac{\pi^{\text{split}}_{C|C^\star,K^\star}(C|C^\star,K^\star)}{\pi^{\text{merge}}_{C^\star|C,K}(C^\star|C,K)} = \frac{\binom{|k|}{|k_1|}\frac{1}{|k|-1}\frac{1}{\sum_{i=1}^{K}\mathbb{1}_{|i|>1}}}{\binom{K}{2}}$$

De plus, $\frac{\pi^{\mathbf{N}}_{theta|C,K}(\theta|C,K,\mathbf{N})}{\pi^{\mathbf{N}}_{\theta|C,K}(\theta^\star|C^\star,K^\star,\mathbf{N})}$ est le ratio des distributions marginales *a posteriori* des paramètres θ et est égal à

$$\frac{\pi_\theta^{\mathbf{N}_k}(\theta_k|\mathbf{N}_k)}{\pi_\theta^{\mathbf{N}_{k_1}}(\theta_{k_1}|\mathbf{N}_{k_1})\pi_\theta^{\mathbf{N}_{k_2}}(\theta_{k_2}|\mathbf{N}_{k_2})}$$

où $\mathbf{N}_k$ correspond au nombre d'individus dans le groupe k. $\pi_\theta^{\mathbf{N}_k}(\theta|V_k)$ est divisé comme suit :

$$\pi_\theta^{\mathbf{N}_k}(\theta|\mathbf{N}_k) = \prod_l^L \pi^{\mathbf{N}_k}_{pqm|l,k}(p_l,q_l,m_l|\mathbf{N}_k)\ \pi^{\mathbf{N}_k}_{d|k}(d|\mathbf{N}_k)\ \pi^{\mathbf{N}_k}_{f|k}(f|\mathbf{N}_k))$$

où

$$\pi^{\mathbf{N}_k}_{pqm|l,k} \equiv \mathcal{D}(1 + n_{llk}, 1 + n_{l(l+1)k}, 1 + n_{l\dagger k})$$

n_{llk}, $n_{l(l+1)k}$ et $n_{l\dagger k}$ correspondent au nombre d'individus dans le groupe k qui restent dans la classe l, passent de la classe l à classe $l+1$ ou meurent. De plus, la distribution *a posteriori* de la structure diamétrique d est égale à

$$\pi^{\mathbf{N}_k}_{d|k} \equiv \mathcal{D}(1 + n_{lk}, \ldots, 1 + n_{Lk})$$

où n_{lk} est le nombre d'individus dans le groupe k dans la classe l à l'état initial t. Enfin le taux de recrutement est donné,

$$\pi^{\mathbf{N}_k}_{f|k} \equiv \mathcal{G}\left(0.01 + n_{01k}, \frac{1}{n_k + 1}\right)$$

où n_k est le nombre total d'individus dans le groupe k à l'état initial t et n_{01k} le nombre de recrutés dans le groupe k.

2.3.3 Conclusions et perspectives

L'algorithme proposé permet simultanément de grouper les espèces, estimer les paramètres et trouver le nombre de groupes dans le cadre d'un modèle de dynamique des populations. L'approche bayésienne offre des avantages en particulier celle de la prise en compte d'information *a priori*. Par exemple, il est possible de supposer que le nombre de recrutés par espèce correspond à 1% de la population de l'espèce ou encore d'interpréter la loi *a priori* du nombre de groupes selon la théorie des niches ou de la théorie neutre (Hubbell, 2001). Néanmoins, cette approche se heurte à certains problèmes. On peut citer en particulier celui qui consiste à construire un algorithme efficace qui permette de visiter un ensemble important de configurations ou encore d'incorporer des covariables pour modéliser les probabilités de transitions. Ce dernier point est abordé dans la section suivante.

2.3.4 Modèles de Usher inhomogènes en mélange (Ouédraogo et al., 2013; Mortier et al., 2015)

Le modèle précédent présente l'avantage de modéliser simultanément l'ensemble des espèces en particulier les espèces peu abondantes mais présente l'inconvénient (i) de ne pas pendre en compte l'environnement biotique et abiotique dans le calcul des probabilités, (ii) de regrouper les espèces dans les mêmes groupes pour l'ensemble des processus et (iii) d'utiliser, pour estimer les probabilités de transition, les informations par classes de diamètre et non l'information disponible de la croissance individuelle (Picard et al.,

2012). Pour y remédier, nous avons récemment proposer de développer un modèle de Usher non homogène en mélange et dénommé MIMM (« Mixture of Inhomogeneous Matrix Models »). L'approche utilisée combine des méthodes de régression en mélange avec sélection de variable par maximum de vraisemblance pénalisée.

Modèle

Tout comme dans le travail précédent, nous supposons que la dynamique de l'espèce s est modélisée par un modèle de Usher. L'équation 2.10 peut se réécrire de la façon suivante :

$$\mathbb{E}(\mathbf{N}_{s,t+1}|\mathbf{N}_{s,t}) = \mathbf{P}^{\bullet}_{s,t}\mathbf{S}_{s,t}\mathbf{N}_{s,t} + \mathbf{R}_{s,t}$$

où

$$\mathbf{P}^{\bullet}_{s,t} = \begin{pmatrix} 1 - q^{\bullet}_{s,2,t} & & 0 & 0 \\ q^{\bullet}_{s,2,t} & \ddots & & \vdots \\ & \ddots & 1 - q^{\bullet}_{s,L,t} & 0 \\ 0 & & q^{\bullet}_{s,L,t} & 1 \end{pmatrix}$$

correspond à la matrice de transition sachant que les arbres sont vivants,

$$\mathbf{S}_{s_t} = \begin{pmatrix} 1 - m_{s,1,t} & & 0 \\ & \ddots & \\ 0 & & 1 - m_{s,L,t} \end{pmatrix}$$

est la matrice de survie et enfin

$$\mathbf{R}_{s,t} = \begin{pmatrix} r_{s,t} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

est le vecteur de recrutement. À la différence du modèle précédent (voir eq. 2.11), nous supposons que les probabilités de transition dépendent du temps. Comme l'objectif est de prendre en compte l'environnement pour modéliser la dynamique forestière, nous avons supposé que ces probabilités dépendent du temps au travers de l'environnement, de sorte que

$$q^{\bullet}_{s,l,t} \equiv q^{\bullet}_{s,l}(X^{G}_{s,l,t}), \quad m_{s,t,1} \equiv m_s(X^{M}_{s,l,t}) \quad \text{et} \quad r_{s,t} \equiv r_s(X^{R}_{s,t})$$

où X^{G}_{t}, X^{M}_{t}, et X^{R}_{t} correspondent à un ensemble de covariables mesurées au temps t associée aux processus de croissance, mortalité et recrutement.

D'un point de vue pratique, pour comprendre la dynamique forestière, nous disposons du suivi annuel de la croissance individuelle des arbres, de la mortalité/survie de chaque arbre et des nouveaux recrutés. Cette information permet de considérer les estimateurs par régression (Rogers-Bennett and Rogers, 2006; Picard et al., 2008b). Ainsi les prédicteurs des taux de croissance, de la survie et du recrutement peuvent s'exprimer de la façon suivante :

- pour la croissance

$$q^{\bullet}_{s,l+1}(x^G_t) = \frac{a_{s,l}(x^G_{s,l,t})}{d_l}$$

où $a_{s,l}(x^G_{s,l,t})$ est le taux de croissance « typique » de la classe l, d_l correspond au diamètre de la classe l, et

$$a_{s,l}(x^G_{s,l,t}) = x^G_{s,l,t}\beta_s$$

- pour la mortalité

$$m_{s,l}(x^M_{s,l,t}) = \text{logit}^{-1}\left(X^M_{s,l,t}\gamma_s\right)$$

- pour le recrutement

$$r_{s,t} = \exp\left(X^R_{s,t}\alpha_s\right)$$

β, γ et α sont des paramètres à estimer. Pour cela nous supposons de simples modèles linéaires généralisés :

- Soit ΔD_{stj} l'accroissement diamétrique de l'individu j de l'espèce s entre les temps t et $t+1$,

$$\Delta D_{stj} = \mu^G_0 + log(D_{stj})\mu^G_1 + D_{stj}\mu^G_2 + X^G_{stj}\beta_s + \varepsilon_s$$

où μ^G's et β_s's sont les paramètres inconnus,

- En posant M_{stj} la mortalité de l'individu j de l'espèce s entre les temps t et $t+1$, alors

$$\begin{aligned} M_{stj} &\sim \mathcal{B}er\left(m_{stj}\right) \\ \text{logit}(m_{stj}) &= \mu^M_0 + \log(D_{stj})\mu^M_1 + D_{stj}\mu^M_2 + X^M_{tj}\gamma_s \end{aligned}$$

où μ^M's et γ_s's sont à déterminer,

- et enfin si N_{st} désigne le nombre de recrutés de l'espèce s entre les temps t et $t+1$, :

$$\begin{aligned} R_{st} &\sim \mathcal{P}\left(r_{st}\right) \\ \log(r_{st}) &= \mu^M_s + X^R_{st}\alpha_s \end{aligned}$$

où μ^M et α_s les paramètres.

Mais en raison de la richesse spécifique, le nombre d'observations pour une espèce donnée peut être insuffisant pour estimer de façon correcte les paramètres de ces différents modèles. A nouveau, nous proposons d'utiliser les modèles de mélange pour surmonter ces difficultés. Nous proposons ainsi de regrouper les espèces ayant des caractéristiques similaires pour chaque processus en réponse à l'environnement. Le cadre formel est donc les modèles de régression en mélange. Si on note $\boldsymbol{\psi}$ le vecteur des paramètres pour chaque processus, les fonctions de log-vraisemblance associées s'expriment alors

$$\ell(\boldsymbol{\psi}|\mathbf{Y}) = \sum_{s=1}^{S} \log \left[\sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} \prod_{j=1}^{n_{st}} f\left(Y_{stj}|X, \boldsymbol{\psi}_k\right) \right] \tag{2.16}$$

avec f la fonction de densité gaussienne et $Y_{sti} = \Delta D_{sti}$ dans le cas du processus de croissance, ou avec f la fonction de probabilité de la loi de Bernoulli et $Y_{sti} = M_{sti}$ le processus de mortalité et enfin dans le cas du recrutement :

$$\ell(\boldsymbol{\psi}|\mathbf{Y}) = \sum_{s=1}^{S} \log \left[\sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} f\left(R_{st}|X, \boldsymbol{\psi}_k\right) \right] \tag{2.17}$$

où f est la fonction de masse d'une loi de Poisson.

Il est important de noter que l'on ne cherche pas à regrouper les observations, mais les espèces. De plus, nous pouvons supposer que les variables environnementales agissent différemment sur chaque groupe d'espèces. Nous souhaitions donc sélectionner les covariables pertinentes tout en regroupant les espèces dans des groupes homogènes. Les paramètres présents dans les équations 2.16 et 2.17 s'obtiennent en maximisant les vraisemblances pénalisées suivantes :

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \left\{ \ell(\boldsymbol{\psi}|\mathbf{Y}) - p_n(\boldsymbol{\psi}) \right\} \tag{2.18}$$

où p_n est un terme de pénalité. Dans ce travail nous utilisons la pénalisation appelée *LASSO adaptative* proposée par Zou (2006) :

$$p_n(\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \eta_{nk} \sum_{q=1}^{Q} \frac{|\psi_{kq}|}{|\hat{\psi}_{kq}|}$$

avec ψ_{kq} le $q^{\text{ème}}$ élément de $\boldsymbol{\psi}_k$, $\hat{\psi}_{kq}$ désigne le maximum de vraisemblance de ψ_{kq} et η_{nk} est un paramètre calibré par validation croisée. Ce travail généralise les travaux de Khalili and Chen (2007)et Städler et al. (2010) au cas multivarié. Supposons maintenant que l'on ait obtenu K_g groupes de croissance, K_r groupes de recrutement et K_m groupes de mortalité, en croisant les différentes classifications nous obtenons $K_g \times K_r \times K_m$ combinaisons

de groupes, appelé $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$. Chaque espèce appartient à un unique $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$ groupe. Ainsi on obtient ce que l'on a appelé le mélange de modèle de Usher inhomogène :

$$\mathbf{N}_{s,t+1} = \mathbf{P}^{k_G}(X_t)\mathbf{S}^{k_M}(X_t)\mathbf{N}_{s,t} + \mathbf{R}^{k_R}(X_t) \tag{2.19}$$

Algorithme Expectation-Maximization (EM)

Pour estimer les paramètres et l'assignation des espèces aux groupes nous avons adapté un algorithme EM. La log-vraisemblance complétée s'écrit

$$\ell_c(\boldsymbol{\psi}|\mathbf{Y},\mathbf{Z}) = \sum_{s=1}^{S}\sum_{k}^{K}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} z_{sk}\log(f(y_{stj}|\psi_k)) + \sum_{s=1}^{S}\sum_{k}^{K} z_{sk}\log\pi_k \tag{2.20}$$

où $\mathbf{Z}$ désigne le vecteur latent de classe.

L'étape E consiste à calculer l'espérance la log-vraisemblance complétée 2.20. Cela s'obtient aisément et permet d'obtenir, à l'itération $m+1$, la probabilité que l'espèce s appartienne au groupe k :

$$w_{kstj}^{(m+1)} = w_{ks}^{(m+1)} = \frac{\pi_k^{(m)}\prod_{t'=1}^{T}\prod_{j'=1}^{n_{st'}} f(\mathbf{Y}_{st'j'}|\mathbf{X},\boldsymbol{\psi}_k^{(m)})}{\sum_{l=1}^{K}\pi_l^{(m)}\prod_{t'=1}^{T}\prod_{j'=1}^{n_{st'}} f(Y_{st'j'}|\mathbf{X},\boldsymbol{\psi}_l^{(m)})}$$

L'étape M consiste à maximiser l'espérance des log-vraisemblances complétées. Pour ce qui est de la mise à jour des π_k nous avons adopté l'approche empirique proposée par Khalili and Chen (2007) :

$$\pi_k^{(m+1)} = \frac{1}{S}\sum_{s=1}^{S} w_{ks}^{(m+1)}.$$

Cette équation est une approximation. En effet, les π_k interviennent dans la pénalisation. Néanmoins, les simulations ont démontré que cette approximation fonctionnait bien. Städler et al. (2010) discutent d'une approche générale mais qui est plus délicate à mettre en pratique. Enfin, les paramètres sont solutions des équations suivantes :

1. pour le processus de croissance

$$\hat{\beta}_k^{(m+1)} = \arg\max_{\beta_k}\left\{\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} w_{stjk}^{(m+1)}\log f\left(\Delta D_{stj}|X_{kj}^{G}\beta_k,\sigma_k^2\right) - \pi_k^{(m+1)}\eta_{nk}\frac{|\beta_k|}{|\hat{\beta}_k|}\right\} \tag{2.21}$$

   où f est la densité gaussienne.

2. pour le processus de mortalité

$$\hat{\gamma}_k^{(m+1)} = \arg\max_{\gamma_k} \left\{ \sum_{s=1}^{S} \sum_{t=1}^{T} \sum_{j=1}^{n_{st}} w_{stjk}^{(m+1)} \log f\left(M_{stj} | X_{kj}^{M} \gamma_k\right) - \pi_k^{(m+1)} \eta_{nk} \frac{|\gamma_k|}{|\hat{\gamma}_k|} \right\} \tag{2.22}$$

où f correspond à la fonction de masse d'une loi de Bernoulli.

3. pour le processus de recrutement

$$\hat{\alpha}_k^{(m+1)} = \arg\max_{\alpha_k} \left\{ \sum_{s=1}^{S} \sum_{t=1}^{T} w_{stk}^{(m+1)} \log f\left(R_{st} | X_k^{R} \alpha_k\right) - \pi_k^{(m+1)} \eta_{nk} \frac{|\alpha_k|}{|\hat{\alpha}_k|} \right\} \tag{2.23}$$

où f est la fonction de masse d'une loi de Poisson.

La mise en œuvre de cet algorithme s'est avérée beaucoup plus simple que prévu. En effet nous avons pu intégrer les fonctionnalités du package `glmnet` (Friedman et al., 2010) au package `flexmix` (Leisch, 2004; Grün and Leisch, 2007, 2008). Nous avons proposé aux auteurs de ce dernier package le code et la fonction dénommée `FLXMRglmnet` a alors été intégrée au package `flexmix`.

2.3.5 Conclusions et perspectives

Cette approche présente l'avantage de combiner dans un cadre théorique des méthodes sophistiquées et modernes pour la recherche de covariables dans le cadre des modèles linéaires généralisés, tout en étant simple d'usage pour ce qui concerne la dynamique des peuplements forestiers. Mais le modèle que je propose présente lui aussi quelques lacunes, en particulier, celle de supposer que les observations réalisées au cours du temps sur un même individu ou entre les individus d'une même espèce sont indépendantes. L'utilisation des effets aléatoires est une piste intéressante pour surmonter cette difficulté. Différents auteurs ont déjà proposé des méthodes de sélection de variables par vraisemblance pénalisée dans le cadre des modèles linéaires généralisés à effets mixtes (Chelldorfer et al., 2011; Schelldorfer et al., 2013; Groll, 2015). Le travail que mène Romain Gaspard pendant sa thèse, thèse que je co-encadre avec Bruno Hérault, consiste à étendre les résultats obtenus en tenant compte des dépendances entre les mesures réalisées sur un même individus. Il se focalise en particulier sur les modèles de mélange avec effets aléatoires.

2.4 Modèles de distribution des espèces, interprétation et prédiction : la méthode SCGLR (Bry et al., 2013, 2015)

2.4.1 Introduction

Comprendre comment se structurent, dans l'espace, les communautés d'espèces forestières, est un objectif majeur des projets que nous menons (CoForChange, CoForTips), mais aussi des sciences de l'écologie. La masse et la richesse des données collectées au cours des projets a nécessité le développement d'outils statistiques adaptés. Les projets CoForChange et CoForTips offrent une occasion unique de tester ces nouvelles méthodes et en particulier la régression linéaire généralisée sur composantes supervisées (SGLR). La régression dans le contexte des modèles linéaires généralisés (GLM) est communément employée pour modéliser les distributions d'espèces (Elith and Leathwick, 2009). Or, dans l'estimation courante d'un GLM, la structure de corrélation des régresseurs n'est pas utilisée pour trouver des structures prédictives fortes. La recherche de combinaisons linéaires des régresseurs qui maximisent simplement la vraisemblance du GLM a deux conséquences majeures : (i) la colinéarité des régresseurs est un facteur d'instabilité de l'estimation, (ii) le modèle pouvant s'ajuster à des dimensions de bruit, ses pouvoirs explicatif et prédictif sont fragilisés.

L'idée, développée avec X. Bry et C. Trottier de l'université de Montpellier et G. Cornu du Cirad, est de chercher des structures, appelées composantes, dans l'espace engendré par les covariables, qui à la fois résument l'information contenue dans le tableau des régresseurs et prédisent au mieux q variables d'intérêts ou dépendantes $Y = (Y^1, \ldots, Y^q)$. Les composantes devront de plus être orthogonales entre elles pour éviter les redondances. La figure 2.2 résume graphiquement le problème et la méthode « SCGLR". Il est important de noter que les composantes seront les mêmes pour toutes les variables à expliquer $Y^j, j = 1, \ldots, q$ mais que leurs effets seront propres à chacune d'elle (Bry et al., 2013, 2015).

2.4.2 Modèle

Considérons la situation suivante où chaque $Y_i^j, j = 1, \ldots, q$ et $i = 1, \ldots, n$ est une variable aléatoire dont la loi appartient à la famille exponentielle (chaque variable Y^j peut avoir sa propre loi). On note $x_i = (x_i^1, \ldots, x_i^p)$ les p covariables observées pour chaque i et supposées fixes et connues. L'idée de la méthode repose avant tout sur la reformulation de la régression dans le

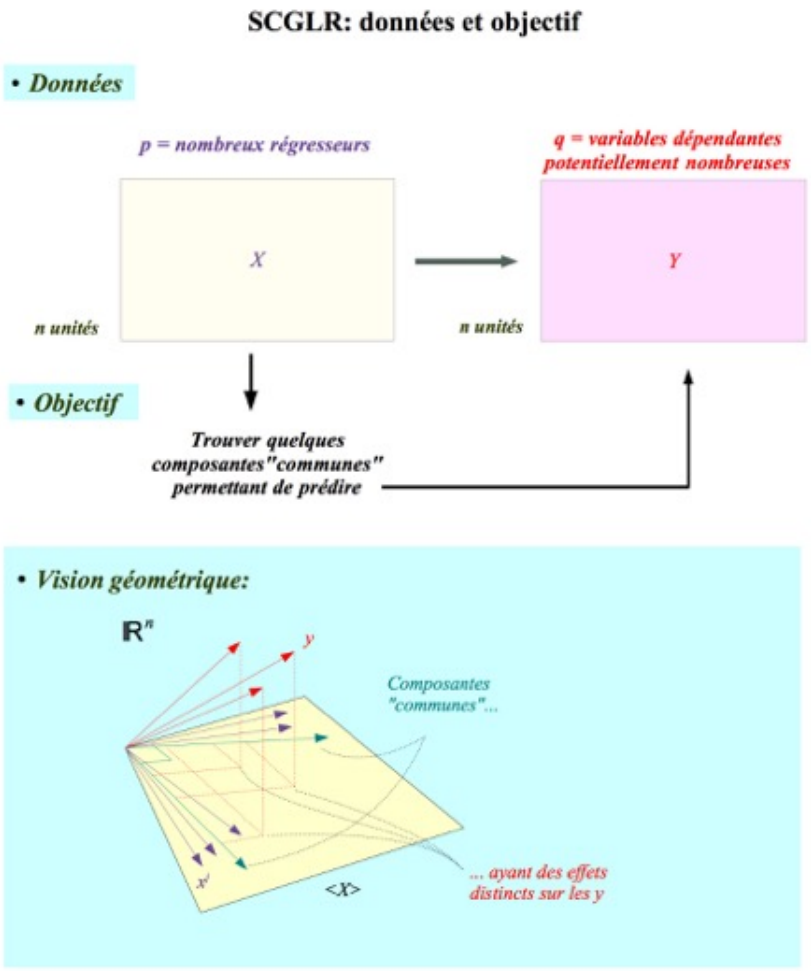


FIGURE 2.2: Présentation graphique de la méthode « SCGLR ».

cadre des modèles linéaires généralisés

$$\begin{aligned} Y^j &\sim f(y, \eta^j) \\ \eta^j &= X\beta^j \end{aligned}$$

de sorte que le prédicteur linéaire η^j se réécrive de la façon suivante :

$$\eta^j = Xu\gamma^j$$

où u est un vecteur de dimension p de norme 1 et $\gamma^j, j = 1, \ldots, q$ les coefficients associés. u est appelé classiquement vecteur des scores ou « loadings » et représente la direction d'intérêt dans l'espace engendré par X. Cette réécriture ne change guère de la simple régression et la rend à première vue plus compliquée puisque l'estimation n'est plus directe. En effet, le produit $u\gamma$ induit naturellement une non-linéarité. Néanmoins, l'utilisation d'algorithmes d'optimisation tels que ceux présents dans le package « nloptr » rend l'estimation aisée (Johnson, 2014). Cette écriture présente en outre de nombreux avantages. Elle permet d'une part de réduire la dimension du problème, rendant ainsi les estimations plus stables, mais aussi de rechercher les u qui contiennent l'information pertinente incluse dans les covariables. Par exemple, si l'on choisissait de réaliser une régression sur composantes principales, u serait immédiatement connu et construit à partir des vecteurs propres issus de l'analyse en composantes principales du tableau X. Dans l'approche « SCGLR », nous proposons d'utiliser l'un des deux critères suivants :

- le critère VC (« Variance Components") égal à

$$\phi(u) = u'X'WXu \tag{2.24}$$

  où W est la matrice des poids associée aux individus et correspond généralement à la matrice diagonale dont les éléments valent $1/n$. Les u maximisant ce seul critère sont alors les scores obtenus par l'analyse en composantes principales du tableau X.
- le critère VPI (« Variable Powered Inertia") qui est égal à :

$$\phi(u) = \left(\sum_{k=1}^{p} \left(u'X'Wx^k x^{k\prime}WXu \right)^l \right)^{\frac{1}{l}} \tag{2.25}$$

  l est un paramètre que l'on fixe. Si $l = 1$, la maximisation du critère VPI permet d'extraire la structure de variance maximale, ce qui correspond à la première composante de l'analyse en composantes principales (ACP). Si l augmente, la maximisation du critère se focalise sur des structures internes à X toujours plus locales comme le montre la figure 2.3.

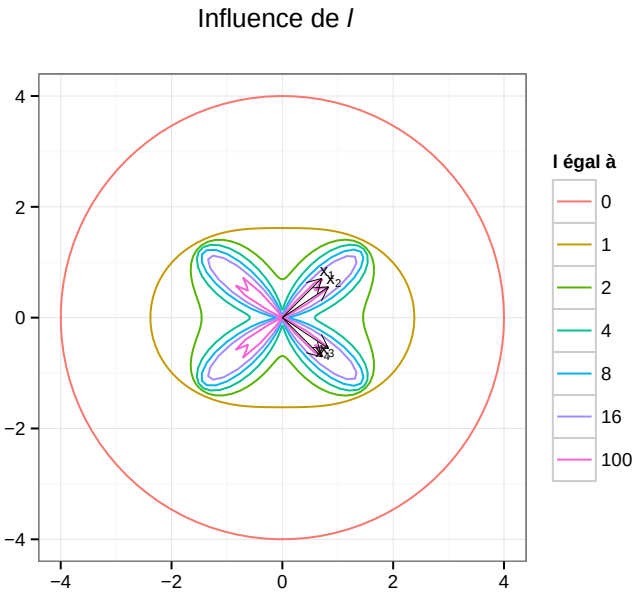


FIGURE 2.3: Courbes d'iso-valeur de VPI (équation 2.25) selon la valeur de l : influence de l sur la localité des faisceaux obtenus.

Mais choisir les u en ne tenant compte que de X n'assure en rien que les vecteurs ainsi obtenus soient adaptés pour prédire au mieux les observations y_i^j. L'originalité de « SCGLR » consiste à chercher les u qui maximisent un compromis entre la fonction ϕ et la qualité d'ajustement du modèle de Y. Le problème peut donc être posé comme celui de l'optimisation du critère suivant :

$$(\hat{u}, \hat{\gamma}) = \arg\max_{u,\gamma} \phi(u)^s \psi(y, u, \gamma)^{1-s} \quad \text{avec} \quad u'u = 1 \tag{2.26}$$

où $\psi(y, u, \gamma)$ est la log-vraisemblance du modèle linéaire généralisé et contient donc l'information relative à la qualité d'ajustement. s est un paramètre de réglage permettant de pondérer l'influence de chaque terme dans le critère global que l'on veut maximiser. Si s vaut zéro, la procédure conduit à maximiser la log-vraisemblance uniquement. En revanche, si s vaut 1 l'algorithme retrouve les directions de maximisation locale de ϕ. L'optimisation de ce critère est réalisé en adaptant l'algorithme des scores de Fisher (FSA) dans lequel la recherche des vecteurs u est obtenue grâce à l'algorithme PING (« Projected Iterated Normalized Gradient »). Les détails sont présentés dans Bry et al. (2015). Le problème ainsi présenté suppose que u est un unique vecteur. Or cela peut s'avérer insuffisant pour prédire correctement les variables dépendantes Y^j. Il est alors utile de chercher d'autres composantes

prédictives. Pour éviter les redondances, les composantes suivantes seront obtenues de sorte à être orthogonales aux précédentes. Le choix du nombre de composantes se fait par validation croisée.

Un « Package » R a été développé pour assurer la distribution de cette méthode auprès d'un large public. Ce package, appelé « SCGLR », est disponible depuis 2013 sur le site du « Comprehensive **R** Archive Network » (CRAN [1]). Nous n'avons cessé de l'améliorer en proposant toujours plus de fonctionnalités. Nous en sommes actuellement à la version 2.1 et de nouvelles modifications sont en cours (Cornu et al., 2015). Ce package contient deux fonctions principales, *scglr* et *scglrCrossVal*, qui permettent respectivement d'estimer les paramètres et de choisir le nombre de composantes par validation croisée. À ce jour, quatre lois de probabilité peuvent être choisies pour modéliser la distribution des Y^j, les lois de Bernoulli, binomiale, de Poisson et normale. Par défaut, le critère VPI (équation 2.25) est utilisé, mais peut être changé si l'utilisateur souhaite utiliser le critère VC (equation 2.24). Différentes méthodes sont aussi proposées : *print*, *summary* ou *plot*. Enfin, un jeu de données test est mis à disposition. Il contient d'une part les données d'abondance de 30 essences forestières mesurées sur 1000 parcelles et d'autre part une cinquantaine de covariables. Celles-ci caractérisent les conditions pluviométrique et topographique de chacune des parcelles ainsi que l'activité photo-synthétique au travers de 23 indices de végétation (Enhanced Vegetation Index, EVI).

2.4.3 Conclusions et perspectives

« SCGLR » présente l'avantage d'être simple à mettre en œuvre et permet d'obtenir des résultats relativement aisés à interpréter. Cette approche s'est avérée particulièrement utile dans la compréhension des liens existant entre les conditions climatiques et la distribution des espèces forestières dans le cadre des projets CoForChange et CoForTips. La mise à disposition d'un package **R** est aussi apparue fort utile, permettant à nos collègues biologistes de s'approprier l'outil et de réaliser eux-mêmes les analyses. Différentes pistes sont d'ores et déjà à l'étude. Dans sa version initiale, la méthode « SCGLR » suppose que les covariables X forment un ensemble unique et homogène. Chaque composante est donc définie en tenant compte de l'ensemble des covariables. Or il est fréquent que X se décompose en un ensemble de thèmes homogènes, par exemple les conditions environnementales (pluviométrie, ensoleillement,...) d'un côté et les caractéristiques photo-synthétiques (EVI) de

1. https://cran.r-project.org/

l'autre. L'idée est de développer une méthode que nous appellerons « Theme-SCGLR » qui consiste à rechercher, dans la même procédure d'estimation, les composantes associées à chaque thème (Bry et al., 2015). Cela favorisera l'interprétation des résultats et des composantes. De plus, « SCGLR » suppose jusqu'ici que les observations sont indépendantes entres elles. Ceci est, dans de nombreuses situations, peu réaliste. L'inclusion d'effets aléatoires est une piste intéressante que Jocelyn Chauvet a abordée pendant son stage de Master en bio-statistiques de l'université de Montpellier, en 2015 et qu'il doit poursuivre lors de son travail de thèse (encadré par Catherine Trottier et Xavier Bry). Enfin, la méthode suppose que les composantes sont les mêmes pour toutes les variables réponses $Y^j,\ j = 1, \ldots, q$. Or cette hypothèse peut être mal appropriée. L'idée est de développer un modèle de mélanges de régression sur composantes supervisées. Le mélange portera sur les variables réponses et non sur les unités statistiques, de manière similaire à ce que j'ai proposé dans le cadre du modèle MIMM (Mortier et al., 2015).

3
Et demain ?

Dans les deux chapitres précédents, j'ai tenté de présenter succinctement le contexte biologique et les outils statistiques sur lesquels je me suis appuyé pour développer mes recherches, tout en essayant de les illustrer par quelques résultats clés. Ces recherches sont-elles pour autant closes ? Bien évidemment non. D'abord, toutes les recherches que j'ai jusqu'alors menées présentent un certain nombre de lacunes - hypothèses d'indépendance, environnement biotique et abiotique insuffisamment pris en compte - mais surtout les écosystèmes, outre qu'ils n'ont pas livré tous leurs secrets, sont loin d'être à l'abri des menaces extérieures.

Pour avancer, il me semble nécessaire de définir une stratégie de recherche qui ne peut être menée à bien que grâce à un dispositif matériel et humain - direction d'étudiants en thèse ou en post-doctorat, mécanisme de formation, nouveaux partenariats, réseaux national et international - permettant de fédérer les recherches afin de répondre à l'objectif commun : la préservation des écosystèmes forestiers compatible avec le développement des populations humaines.

3.1 Stratégie de recherche

Les thèmes de recherche que je souhaite aborder, dans les années à venir, doivent toujours combiner les deux points de vue centraux de mon métier (i) une recherche appliquée pour répondre aux grands enjeux actuels de l'équipe,

du Cirad mais surtout des pays en développement (ii) une recherche théorique pour élaborer des outils novateurs en statistiques et modélisations. Aujourd'hui, pour progresser dans notre compréhension des écosystème forestiers, il m'apparaît nécessaire de mieux incorporer l'homme dans la dynamique des changements mais aussi de mieux comprendre les effets du climat. Plusieurs pistes sont actuellement envisagées :

1. Coupler des modèles écologiques, économiques et sociaux. C'est cette approche que je m'efforce, depuis un an, de développer avec le doctorant Florian Claeys ainsi qu'avec Sylvie Gourlet-Fleury, Alain Karsenty et Philippe Delacotte. L'équipe se focalise pour le moment sur le volet écologique et économique des concessions forestières. Ces dernières jouent un rôle central dans la vie économique et sociale au sein des pays du bassin du Congo. L'idée consiste à coupler les modèles matriciels inhomogènes en mélange que nous avons développés pour modéliser la dynamique écologique des écosystèmes forestiers (Mortier et al., 2015), avec un modèle économique qui traduit les activités des concessions forestières. Cela permettra de quantifier l'effet des instruments économiques internationaux tels que les projets carbone (REDD+) sur les pratiques des exploitants forestiers en tenant compte de différents scénarios climatiques.
2. Faire la part entre ce qui est héritage du passé et effets du climat sur la répartition des espèces. Ceci est particulièrement crucial pour prédire l'influence des changements climatiques sur la végétation. En effet, si les relations entre le climat et la répartition des espèces ne coïncident qu'avec la structure spatiale des conditions environnementales, il est illusoire d'essayer de prédire une quelconque répartition dans le cas où le climat changerait de deux, trois voire huit degrés. L'une des idées consiste à simuler la répartition spatiale des conditions environnementales en utilisant des approches issues de la géo-statistique multivariée et les modèles de co-régionalisation et ce, dans le but d'élaborer un modèle « nul » spatialement explicite et ainsi tester l'effet de l'environnement. Ce travail est imaginé et conçu avec deux collègues, Maxime Réjou-Méchain de l'IRD (anciennement post-doctorant) et Nicolas Desassis de l'école des Mines de Paris.
3. Modéliser la dynamique de l'utilisation des sols. Ceci me semble une étape clé pour proposer des modes de gestion du territoire dans le contexte des changements globaux. Nous travaillons activement, avec Nicolas Bousquet d'EDF (Recherche et Développement) et Nicolas Desassis de l'école des Mines de Paris, à une première ébauche de ce type de modèle. L'idée repose sur le principe suivant :

(a) On observe, depuis un certain nombre d'années, la variation du pourcentage de surface des sols dédiée à la forêt, à l'agriculture et aux jachères (entre autres). On suppose que cette dynamique peut être représentée par un modèle multivarié auto-régressif.

(b) On suppose de plus que, d'un commun accord, les pays d'Afrique centrale souhaitent à l'objectif 2050, conserver globalement un certain pourcentage de terre dédié aux forêts, mais aussi développer l'agriculture tout en conservant un certain pourcentage de jachères.

(c) Connaissant ces objectifs, il est alors possible d'estimer les efforts que chacun devrait consentir pour les atteindre et de simuler les résistances aux changements et les incitations extérieures susceptibles d'influer sur les pratiques.

D'un point de vue statistique, j'envisage différentes thématiques qui m'apparaissent porteuses et innovantes. Deux points doivent impérativement être étudiés et concernent :

1. la généralisation du modèle MIMM (Mortier et al., 2015) qui a démontré une certaine efficacité quant à la qualité de prédiction de la dynamique forestière. J'entends, en particulier, me focaliser sur deux aspects :
   (a) la prise en compte des dépendances entre les mesures. Cela conduit naturellement à s'interroger sur les modèles de mélange pour données longitudinales. J'aborde cette question avec différents collaborateurs, Marie Denis, Bruno Hérault, Sylvie Gourlet-Fleury et Mahlet Tadesse,
   (b) la sélection de variables dans le contexte des modèles de mélange pour des données groupées.
2. la généralisation de l'approche « SCGLR » (Bry et al., 2013, 2015) pour des données non indépendantes. Deux points sont *a minima* cruciaux à étudier :
   (a) les résultats théoriques,
   (b) la prise en compte des dépendances spatiales, temporelles et spatio-temporelles.

Un autre point, plus novateur au vu de mes activités passées, porte sur la prise en compte

1. des liens non-linéaires entre les processus d'intérêts (croissance, recrutement, mortalité) et les variables environnementales ;
2. des caractéristiques fonctionnelles des processus étudiés.

L'étude de la totalité de ces questions relève d'un travail collectif. Elle nécessite entre autres choses, collaborations et contributions d'étudiants.

3.2 L'encadrement

Les étudiants en thèse et en post-doctorat jouent, dans l'optique définie plus haut, un rôle déterminant. Aujourd'hui, je participe, comme je l'ai indiqué dans la première partie de ce document, au co-encadrement direct de trois étudiants et à la co-direction de deux nouvelles thèses récemment inscrites.

Le co-encadrement, la co-direction de thèses en cours

- la thèse de Romain Gaspard, débutée en 2014 et que je co-dirige avec Bruno Hérault, Sylvie Gourlet-Fleury et Mahlet Tadesse porte sur la question de la résilience des forêts tropicales humides aux effets combinés de l'exploitation forestière et des changements climatiques. Ce travail revêt, en particulier, deux aspects méthodologiques :
  1. Le développement de modèles mixtes en mélanges pour modéliser les processus de croissance, mortalité et recrutement en tenant compte des dépendances entre les mesures faites sur un même individu ou des individus d'une même espèce. Ceci est une première généralisation du travail publié récemment (Mortier et al., 2015).
  2. Le couplage de modèle de croissance (gaussien ou log-gaussien) et des modèles de survie (Cox ou Bernoulli) en tenant compte de la diversité spécifique (modèle de mélange et à effets aléatoire).
- La thèse de Florian Claeys, commencée en 2013 sous la direction d'Alain Karsenty, Philippe Delacote et Sylvie Gourlet-Fleury, a pour objectif d'étudier l'impact de différents instruments internationaux de types REDD+ sur l'évolution des pratiques des exploitants forestiers. La spécificité de ce sujet repose sur
  1. la combinaison du modèle MIMM et d'un modèle économique d'exploitation forestière,
  2. l'utilisation de scénarios d'exploitation et de changements climatiques pour quantifier et prédire l'état des forêts,
  3. la simulation des effets d'outils économiques internationaux sur les dynamiques forestières.

– Le post-doctorat de Jean-François Bastin qui travaille avec Raphaël Pellissier et Sylvie Gourlet-Fleury dans le cadre du projet CoForTips pour définir une méthode de détection de rupture au sein des forêts tropicales et mettre ainsi en avant les caractéristiques fonctionnelles de ces écosystèmes sous pression anthropique et changements climatiques. Ce travail repose à nouveau sur l'utilisation des modèles de mélange. La spécificité ici tient au fait que

  1. la réponse est multivariée et appartient au simplex. Chaque parcelle observée est composée d'un pourcentage de pionnières/décidues, pionnières/sempervirentes, tolérantes à l'ombre/décidues ou encore tolérantes à l'ombre/sempervirentes, la somme de ces pourcentage valant 100.
  2. les variables dépendantes, tout comme les probabilités *a priori* π_k d'appartenir au groupe k, dépendent de covariables environnementales et anthropiques.

Thèses à venir :

– Alexandra Jestin vient de débuter sa thèse à l'automne 2015. Celle-ci sera co-encadrée par Marie Denis et Mahlet Tadesse. Mlle Jestin travaillera sur la modélisation des processus de croissance et de mortalité en lien avec l'environnement. La spécificité de ce travail repose sur le fait que

  1. les liens entre la réponse et l'environnement seront considérés non linéaires (modèles additifs généralisés, GAM),
  2. les observations ne sont pas indépendantes (modèles additifs généralisés à effets aléatoires),
  3. les espèces répondent différemment (modèles additifs généralisés à effets aléatoires en mélange),
  4. les covariables peuvent agir de manières spécifiques au cours du temps (sélection dans des modèles additifs généralisés à effets aléatoires en mélange).

– la thèse de Jocelyn Chauvet qui sera encadrée principalement par Catherie Trottier et Xavier Bry devrait permettre de généraliser les résultats obtenus dans le cadre de « SCGLR » et de les étendre au contexte spatial et spatio-temporel.

L'ensemble de ces travaux est le fruit des recherches et des réflexions que j'ai menées pendant les 13 dernières années ainsi que des collaborations que j'ai construites au fil du temps. Ces travaux restent cohérents entre eux, les uns permettant d'étendre et généraliser des modèles déjà développés, les

autres permettant de combiner ces modèles pour améliorer les résultats déjà obtenus. Ils devraient non seulement permettre de lever certaines hypothèses - indépendance, liens linéaires - mais aussi de mieux incorporer l'homme et les impacts anthropiques dans les modèles.

3.3 La formation

Bien que l'encadrement d'étudiant fasse partie intégrante de la formation, la participation à des cursus d'ingénieur, de Master ou licence reste aussi une priorité en particulier dans les pays du sud mais sans que pour autant il faille négliger ceux du nord. L'équipe réfléchit actuellement à l'élaboration d'un cursus spécifique destiné aux futurs ingénieurs forestiers des pays du bassin du Congo. Ce cursus aurait pour objectif, outre d'enseigner l'écologie ou l'économie et les sciences forestières, de proposer des modules en statistiques et informatiques. La demande est forte et tout doit encore être construit pour y répondre efficacement. Quels que soient les solutions envisagées, il me semble que tout repose sur les partenariats existants et rien ne pourra se faire sans le soutien des organismes de formations traditionnelles, universités et écoles africaines, françaises ou internationales. Différents réseaux existent d'ores et déjà et la mise en place de cette formation ne peut reposer que sur des partenariats et collaborations. Ce qui m'amène à présenter cette dernière partie.

3.4 Le partenariat

Développer des collaborations est, me semble-t-il, une étape nécessaire et cruciale pour élaborer des projets, favoriser des recherches innovantes, pérenniser des systèmes de formation et obtenir les financements nécessaires.

Au niveau national, les collaborations que j'ai pu développer avec l'université de Montpellier, avec l'INRA d'Avignon ou de Bordeaux, avec les membres de l'équipe MORSE d'AgroParisTech, avec l'université Lille I ainsi qu'avec le groupe » Statistiques pour l'environnement » de la Société Française de la Statistique ou encore le groupe de recherche (GDR) dédié à l'écologie statistique me permet de promouvoir les statistiques pour une gestion durable des forêts tropicales. L'existence de ce réseau devrait favoriser l'émergence de projets de recherche et aider à trouver des appuis précieux pour promouvoir la formation des statistiques dans les pays du sud et en particulier dans les pays francophones du bassin du Congo.

Au niveau international, mes partenariats restent encore limités. Il me

semble donc nécessaire de les élargir, en particulier avec les États-Unis en m'appuyant sur la collaboration que j'ai établie depuis quelques années avec les Pr. Tadesse et A. Arab de Georgetown University. Cette coopération m'a offert l'opportunité, depuis 2013, d'être invité à présenter mes travaux d'une part dans quatre universités américaines différentes (en 2013 dans le département de mathématiques et statistiques de Georgetown University (M. Tadesse), en 2014 dans le département de statistiques de Rice University (M. Vanucci), en 2015 dans le département de statistiques de Georges Washington University (T. Apanasovich) et en 2015 dans le département d'entomologie de Madison-Wisconsin University (J. Zhu)) et d'autre part à trois conférences internationales (en 2014 et 2015 dans la conférence « Eastern North America Region » et en 2015 à la réunion annuelle de l'« American Mathematical Society »). De plus, cette collaboration a donné lieu à un premier projet financé par le « Georgetown Environmetal Initiaves » à hauteur de 17,000$. Celui-ci nous a permis d'encadrer deux étudiants de Master. Je m'efforce aussi depuis peu à mettre sur pied un groupe de recherche autour du thème « *Common efforts for tropical forests* ». Cette démarche pourrait alors aisément rejoindre des projets américains ambitieux existants tel que le « Congo Basin Institute » qui visent aussi à promouvoir la formation et la recherche au sein des pays africains. En ce sens, je pense qu'il serait intéressant, pour moi, pour l'équipe et plus globalement pour le Cirad que je sois positionné, pour un temps, à Georgetown University.

Bibliographie

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 :716–723. 35

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, 88 :669–679. 41

Atchade, Y. (2006). An adaptive version for the Metropolis Adjusted Langevin Algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.*, 8 :235–254. 43

Atchade, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11 :815–828. 43

Baillargeon, S. (2005). Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations. Mémoire de maîtrise, Université Laval, Québec. 32

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data*, volume 101 of *Monographs on Statistics & Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL. 31, 32, 46

Bar-Hen, A. and Mortier, F. (2004). Influence and sensitivity measures in correspondence analysis. *Statistics*, 38(3) :207–215. 41

Barbault, R. and Weber, J. (2010). *La vie, quelle entreprise !* Seuil. 21

Barry, R. and Ver Hoef, J. (1996). Blackbox kriging : spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics*, 1 :297–322. 33

Bartholomé, J., Salmon, F., Vigneron, P., Bouvet, J., Plomion, C., and Gion, J. (2013). Plasticity of primary and secondary growth dynamics in eucalyptus hybrids : A quantitative genetics and qtl mapping perspective. *BMC Plant Biology*, 13. 28

Bedel, F., Durrieu de Madron, L., Dupuy, B., Favrichon, V., Maître, H., Bar-Hen, A., and Narboni, P. (1998). Dynamique de croissance dans des peuplements exploités et éclaircis de forêt dense africaine. le dispositif de M'baïki en république centrafricaine (1982-1995). *CIRAD Forêt, Montpellier. Série FORAFRI, document*, 1 :71. 22

Bellwood, D. and Wainwright, P. (2001). Locomotion in labrid fishes : implications for habitat use and cross-shelf biogeography on the great barrier reef. *Coral Reefs*, 20 :139–150. 34

Besag, J. E. (1994). Discussion of paper by U. Grenander and M. I. Miller. *Journal of Royal Statistical Society. Series B*, 56 :591–592. 43

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integratedcompleted likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :719–725. 35

Blaser, J., Sarre, A., and Poore, D. e. a. (2011). Status of tropical forest management 2011. 21

Bliss, C. (1935). The calculation of dosage-mortality curve. *Annals of Applied Biology*, 22 :307–330. 41

Breyer, L. A. and Roberts, G. O. (2000). From Metropolis to diffusions : Gibbs states and optimal scaling. *Stochastic Processes and their Applications*, 90 :181–206. 43

Bry, X., Trottier, C., Mortier, F., Cornu, G., and Verron, T. (2015). Supervised component generalised linear regression with multiple explanatory blocks : Theme-scglr. *PLS 2014*, 1 :1–10. , 39, 61, 64, 66, 69

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component generalized linear regression using a pls-extension of the fisher scoring algorithm. *Journal of Multivariate Analysis*, 119 :47 – 60. , 39, 61, 69

Caswell, H. (2001). *Matrix population models : Construction, analysis and interpretation*. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, second edition. 49

Ceccato, P., Bango, E., Ngouanze, F., , and Damio, T. (1992). étude pédologique des parcelles d'expérimentation des forêts de boukoko et la lolé (M'baïki) (république centrafricaine). 23

Chagneau, P., Mortier, F., and Picard, N. (2009). Designing permanent sample plots by using a spatially hierarchical matrix population model. *Journal Of The Royal Statistical Society Series C*, 58 :345–367. , 34, 40

Chagneau, P., Mortier, F., Picard, N., and Bacro, J. (2011). Prediction of a multivariate spatial random field with continuous, count and ordinal outcomes. *Biometrics*, 58 :345–367. , 32, 34, 40

Chaubert, F., Mortier, F., and André, L. (2008). Multivariate dynamic model for ordinal outcomes. *Journal of Multivariate Analysis*, 99 :1717–1732. 32, 42

Chelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38 :197–214. 60

Chen, M.-H. and Shao, Q.-M. (1999). Properties of prior and posterior distributions for multivariate categorical response data models. *Journal of Multivariate Analysis*, 71 :277–296. 41

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85 :347–361. 41

Christensen, O. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using Generalized Linear Mixed Models. *Biometrics*, 58 :280–286. 32, 43

Christensen, O. F., Møller, J., and Waagepetersen, R. P. (2001). Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalized linear mixed models. *Methodology and Computing in Applied Probability*, 3 :309–327. 43

Clark, J. (2005). Why environmental scientists are becoming bayesians. *Ecology Letters*, 8 :2–14. 29, 31

Cornu, G., Mortier, F., Trottier, C., and Bry, X. (2015). *SCGLR : Supervised Component Generalized Linear Regression*. R package version 2.0.2. 39, 65

Cressie, N. (1991). *Statistics for spatial Data*. John Wiley & Sons. 32

Cressie, N., Caldery, C., Clark, J., Ver Hoef, J., and Wikle, C. (2009). Accounting for uncertainty in ecological analysis : the strengths and limitations of hierarchical statistical modeling. *Agencies and staff of the US department of commerce*. 30, 31

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18 :151–171. 34

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12 :27–36. 47

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). Bayesian variable selection using the gibbs sampler. 38

Diggle, P., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. (With discussion). *Journal of Royal Statistical Society. Series C*, 47 :299–350. 32, 41

Dunstan, P., Foster, S., Hui, F., and Warton, D. (2013). Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, 18 :357–375. 34, 35

Dunstan, P. K., Foster, S. D., and Darnell, R. (2011). Model based grouping of species accross environmental gradient. *Ecological Modeling*, 222 :955–963. 34, 35

Eerikäinen, K., Miina, J., and Valkonen, S. (2007). Models for the regeneration establishment and the development of established seedlings in uneven-aged, norway spruce forest stands of southern finland. *Forest Ecology and Managment*, 242 :444–461. 45

Elith, J. and Leathwick, J. (2009). Species distribution models : ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40 :677–699. 61

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, 7 :509–520. 31

Fan, J. and Jinchi, L. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20 :101–148. 37

FAO (2010). Global forest resources assessment 2010. Forestry Papers. United Nations Food and Agriculture Organization, Rome. 20, 21, 26, 27

Fayolle, A., Engelbrecht, B., Freycon, V., Mortier, F., Swaine, M., RÃ©jouMÃ©chain, M., Doucet, J.-L., Fauvet, N., Cornu, G., and Gourlet-Fleury, S. (2012). Geological substrates shape tree species and trait distributions in African moist forests. *Plos One*, in press. 34

Flores, O., Rossi, V., and F., M. (2009). Autocorrelation offsets zero-inflation in models of tropical saplings density. *Ecological Modeling*, 220 :1797–1809. , 32, 34, 39, 45

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 :1–22. 60

Gaetan, C. and Guyon, X. (2008). *Modélisation et statistique spatiales*. Springer. 33

Garreta, V., Guiot, J., Mortier, F., Chadoeuf, J., and HÃ©ly, C. (2012). Pollen-based climate reconstruction : Calibration of the vegetation-pollen processes. *Ecol. Mod.*, 235-236 :81–94. 32, 34

Gelfand, A. and Ghosh, S. (1998). Model choice : a minimum posterior predictive loss approach. *Biometrika*, 85 :1–11. 31

Gelfand, A. E., Banerjee, S., Sirmans, C. F., Tu, Y., and Ong, S. E. (2007). Multilevel modeling using spatial processes : Application to the Singapore housing market. *Computational Statistics & Data Analysis*, 51 :3567–3579. 40

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient metropolis jumping rules. In Bernardo, J., Berger, J., David, A., and Smith, A., editors, *Bayesian Statistics 5*. Oxford University Press. 31

George, E. and McCulloch, R. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, pages 339–374. 37

Gimenez, O., Buckland, S., Morgan, B., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.-P., Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J. M., Mortier, F., Munoz, F., Ovaskainen, O., Pavoine, S., Pradel, R., Schurr, F., Thomas, L., Thuiller, W., Trenkel, V., de Valpine, P., and Rexstad, E. (2014). Statistical ecology comes of age. *Biology Letters*. 29, 31

Golam Kibria, B., Sun, L., Zidek, J. V., and Le, N. D. (2002). Bayesian spatial prediction of random space-time fields with application to mapping $PM_{2.5}$ exposure. *Journal of the American Statistical Association*, 97 :112–124. 40

Groll, A. (2015). *glmmLasso : Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*. R package version 1.3.5. 60

Grün, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51 :5247–5252. 36, 60

Grün, B. and Leisch, F. (2008). FlexMix version 2 : Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28 :1–35. 36, 60

Grzebyk, M. and Wackernagel, H. (1994). Multivariate analysis and spatial/temporal scales : Real and complex models. In *Proceedings of the XVIIth International Biometric Conference, Hamilton, Ontario*. 33

Guillot, G., Estoup, A., Mortier, F., and Cosson, J. (2005a). A spatial statistical model for landscape genetics. *Genetics*, 170 :1261–1280. 32, 34

Guillot, G., Mortier, F., and Estoup, A. (2005b). Geneland : A program for landscape genetics. *Molecular Ecology Ressources*, 5 :712–715. 32, 34

Guisan, A. and Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecological Modeling*, 135 :147–186. 32

Hoerl, A. and Kennard, R. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12 :55–67. 37

Hubbell, S. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press. 34, 55

Hui, F. C., Warton, D. I., Foster, S. D., and Dunstan, P. K. (2013). To mix or not to mix : comparing the predictive performance of mixture models versus separate species distribution models. *Ecology*, 94 :1913–1919. 34, 35

Hutchinson, G. (1961). The paradox of the plankton. *American Naturalist*, 95 :137–147. 34

Joe, H. (1997). *Multivariate models and dependence concepts*. Monographs on Statistics and Applied Probability. 73. London : Chapman and Hall. xviii, 399 p. . 44

Johnson, S. (2014). The nlopt nonlinear-optimization package. 63

Jones, F. and Muller-Landau, H. (2008). Mesuring long-distance seed dispersal in complex natural environments : an evaluation and integration of classical and genetic methods. *Journal of Ecology*, 96 :642–652. 45

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102 :1025–1038. 58, 59

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52 :119–139. 32

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya : The Indian Journal of Statistics, Series B (1960-2002)*, 60 :65–81. 38

Legendre, P. (1993). Spatial autocorrelation : trouble or new paradigm ? *Ecologia*, 74 :1659–1673. 28

Leisch, F. (2004). FlexMix : A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11 :1–18. 36, 60

Lexerød, N. L. (2005). Recruitment models for different tree species in Norway. *Forests Ecology and Management*, 206 :91–108. 45

Li, Q. and Lin, N. (2010). The Bayesian Elastic Net. *Bayesian Analysis*, 5 :151–170. 38

Lourmas, M. (2003). Diversité génétique et aménagement : utilité d'une modélisation intégrée. *Bois et Forêts des Tropiques*, 276 :85–87. 22

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs &ndash ; a bayesian modelling framework : Concepts, structure, and extensibility. *Statistics and Computing*, 10 :325–337. 31

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative traits.* Sinauer Asociates. Sunderland, 980 p. . 27

Marien, J.-N. and Mallet, B. (2004). Nouvelles perspectives pour les plantations forestières en afrique centrale. *Bois et Forêts des Tropiques*, 282 :67–79. 26

Marin, J.-M. and Robert, C. (2007). *Bayesian core : A pratical approach to computational Bayesian statistics*. Springer text in statistics. Springer-Verlag, New York, NY. 37, 54

Matheron, G. (1963). *Traité de géostatistique appliquée. Tome II : le krigeage.* Mémoires du BRGM, 24. Éditions Bureau de Recherches Géologiques et Minières, Paris. 32

McBratney, A., Odeh, I., Bishop, T., Dunbar, M., and Shatar, T. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma*, 97 :293–327. 40

McCullagh, P. and Nelder, J. (1989). *Generalized linear models.* Monogr. Statist. Appl. Probab. Chapman & Hall/CRC, second edition. 41

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models (Wiley Series in Probability and Statistics).* Wiley-Interscience. 33, 34

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models.* Wiley Series in Probability and Statistics. Wiley. 35

McMahon, S. and Diez, J. (2007). Scales of association : hierarchical linear models and the measurement of ecological systems. *Ecology Letters*, 10 :437–452. 31

Meuwissen, T., Hayes, B., and M.E., G. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157 :1819–1829. 28

Millennium Ecosystem Assessment (2005). *Ecosystems and Human Well-being : Biodiversity Synthesis.* World Resources Institute, Washington, DC. 21

Møller, J. and Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes.* monograph statistics applied probability. Chapman & Hall/CRC, Boca Raton, FL. 43

Mortier, F., Chagneau, P., Etienne, M., Picard, N., Piou, C., and Rossi, V. (2011). Modélisation bayésienne hiérarchique pour l'écologie et la recherche environnementale. 29

Mortier, F., Ouédraogo, D.-Y., Claeys, F., Tadesse, M., Cornu, G., Baya, F., Benedet, F., Freycon, V., Gourlet-Fleury, S., and Picard, N. (2015). Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics*, 26 :39–51. , 36, 39, 48, 55, 66, 68, 69, 70

Mortier, F., Rossi, V., Guillot, G., Gourlet-Fleury, S., and Picard, N. (2013). Population dynamics of species-rich ecosystems : the mixture of matrix population models approach. *Methods in Ecology and Evolution*, 4 :316–326. , 32, 36, 48, 50

Nkoua, M. and Gazull, L. (2013). Les enjeux de la filière "plantations industrielles d'eucalyptus" dans la gestion durable du bassin d'approvisionnement en bois-énergie de la ville de pointe-noire (république du congo). In Farcy, C., Peyron, J.-L., and Poss, Y., editors, *Forêts et foresterie : mutations et décloisonnements.*, pages 175–194, Martinique. Colloque ASRDLF 2011, L'Harmattan. 26

Nobile, A. (2005). Bayesian finite mixtures : a note on prior specification and posterior computation. Technical report, University of Glasgow. 52

Ntzoufras, I., Forster, J., and Dellaportas, P. (2000). Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation*, 68 :23–37. 47

O'Hara, R. and Sillanpää, M. (2009). A review of bayesian variable selection methods : What, how and which. *Bayesian Analysis*, 4 :85–118. 37

Ouédraogo, D.-Y., Mortier, F., Gourlet-Fleury, S., Freycon, V., and Picard, N. (2013). Slow-growing species cope best with drought : evidence from long-term measurements in a tropical semi-deciduous moist forest of central africa. *Journal of Ecology*, 101 :1459–1470. , 36, 48, 55

Parent, E. and Bernier, J. (2007). *Le raisonnement bayésien : Modélisation et inférence*. Springer. 31

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103 :681–686. 38

Picard, N., Bar-Hen, A., Mortier, F., and Chadoeuf, J. (2008a). Understanding the dynamics of an undisturbed tropical rain forest from the spatial pattern of trees. *Journal of Ecology*, 91 :97–108. 34

Picard, N., Bar-Hen, A., Mortier, F., and Chadœuf, J. (2009). The multi-scale marked area-interaction point processes : a model for the spatial pattern of trees. *Scandinavian Journal of Statistics*, 36 :23–41. 34

Picard, N., Köhler, P., Mortier, F., and Gourlet-Fleury, S. (2012). A comparison of five classifications of species into functional groups in tropical forests of French Guiana. *Ecological Modeling*, 11 :75–83. 55

Picard, N., Mortier, F., and Chagneau, P. (2008b). Influence of estimators of the vital rates in the stock recovery rate when using matrix models for tropical rainforests. *Ecological Modelling*, 214 :349–360. 57

Plummer, M. (2003). Jags : A program for analysis of bayesian graphical models using gibbs sampling. 31

Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). In et al., J. B., editor, *Bayesian Statistics 8*, pages 1–45. Oxford University Press. 31

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Asiatic Society. Series B*, 59 :731–792. 35, 52, 53

Roberts, G. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Asiatic Society. Series B*, 60 :255–268. 43

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2 :341–363. 43

Rogers-Bennett, L. and Rogers, D. (2006). A semi-empirical growth estimation method for matrix models of endangered species. *Ecological Modelling*, 195 :237–246. 57

Rue, H., Martino, S., and Chopin, N. (2009). Approxiamte bayesian inference for latent gaussian models using integrated nested laplace approxiamtions. *Journal Of the Royal Statistical Society. Series B*, 71 :319–392. 44

Sagnard, F., Pichot, C., Dreyfus, P., Jordano, P., and Fady, B. (2007). Modelling seed dispersal to predict seedling recruitment : recolonization dynamics in a plantation forest. *Ecological Modeling*, 203 :464–474. 45

Schelldorfer, J., Meier, L., Bühlmann, P., Winterthur, A. X. A., and Zürich, E. T. H. (2013). Glmmlasso : An algorithm for high-dimensional generalized linear mixed models using l1-penalization. *Journal of Computational and Graphical Statistics*. 60

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464. 35

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Asiatic Society. Series B*, 64 :583–639. 31

Städler, N., Bühlmann, P., and Van De Geer, S. (2010). ℓ1-penalization for mixture regression models. *Test*, 19 :209–256. 58, 59

Steneck, R. and Dethier, M. (1994). A functional group approach to the structure of algal-dominated communities. *Oikos*, 69 :DOI : 476–498. 34

Swaine, M. and Whitmore, T. (1988). On the definition of ecological species groups in tropical rain forests. *Vegetation*, 75 :81–86. 34

Tadesse, M., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100 :602–617. 34

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58 :267–288. 36

Usher, M. (1966). A matrix approach to the management of renewable resources, with special reference to selection forests. *Journal of Applied Ecology*, 3 :355–367. 49

Usher, M. (1969). A matrix model for forest management. *Journal of Biometric Society*, 25 :309–315. 49

Vanclay, J. (1992). Modelling regeneration and recruitment in a tropical moist forest. *Canadian Journal of Forest Research*, 22 :1235–1248. 45

Vanclay, J. (1995). Growth models for tropical forests : A synthesis of models and methods. *Forest Science*, 41 :7–42. 48

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92 :1–28. 44

Ver Hoef, J. and Barry, R. (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69 :275–294. 33, 42

Vigneron, P. (1991). Création et amélioration de variétés hybrides d'eucalyptus au congo. pages 345–360. 27

Vigneron, P., Bouvet, J.-M., Gouma, R., Saya, A., Gion, J.-M., and Verhaegen, D. (2000). Eucalypt hybrids breeding in congo. 27

Wackernagel, H. (2003). *Multivariate geostatistics. An introduction with applications.* . Springer, 3rd completely revised edition. 33

Wikle, C. (2003). Hierarchical Bayesian models for predicting the spread of ecology processes. *Ecology*, 84 :1382–1394. 30, 31

Wold, H. (1985). Partial least squares. In Kotz, S. Johnson, N. L., editor, *Encyclopedia of statistical sciences*, pages 581–591. Wiley, New York, 6 edition. 38

Yaglom, A. (1987). *Correlation theory of stationary and related random functions. Volume I : Basic results.* Springer-Verlag, New York. 42

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 :1418–1429. 37, 58

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67 :301–320. 37

Annexes

Autocorrelation offsets zero-inflation in models of tropical saplings density

Ecological Modelling 2013

Ecological Modelling xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Autocorrelation offsets zero-inflation in models of tropical saplings density

O. Flores [a,*], V. Rossi [c], F. Mortier [b]

[a] Centre d'Écologie Fonctionnelle et Évolutive, CNRS – UMR 5175, 1919, route de Mende, 34293 Montpellier Cedex 5, France
[b] CIRAD - UPR Génétique forestière, TA 10/C, Campus international de Baillarguet, 34398 Montpellier Cedex 5, France
[c] CIRAD - UPR Dynamique des forêts naturelles, TA 10/D, Campus international de Baillarguet, 34398 Montpellier Cedex 5, France

ARTICLE INFO

Article history:
Received 10 March 2008
Received in revised form 13 January 2009
Accepted 14 January 2009
Available online xxx

Keywords:
Hierarchical Bayesian Modelling
Conditional Auto-Regressive model
Variable selection
Zero-Inflated Poisson
Posterior predictive
Paracou
French Guiana

ABSTRACT

Modelling the local density of tropical saplings can provide insights into the ecological processes that drive species regeneration and thereby help predict population recovery after disturbance. Yet, few studies have addressed the challenging issues in autocorrelation and zero-inflation of local density. This paper presents Hierarchical Bayesian Modelling (HBM) of sapling density that includes these two features. Special attention is devoted to variable selection, model estimation and comparison.

We developed a Zero-Inflated Poisson (ZIP) model with a latent correlated spatial structure and compared it with non-spatial ZIP and Poisson models that were either autocorrelated (Spatial Generalized Linear Mixed, SGLM) or not (generalized linear models, GLM). In our spatial models, local density autocorrelation was modeled by a Conditional Auto-Regressive (CAR) process. 13 explicative variables described ecological conditions with respect to topography, disturbance, stand structure and intraspecific processes. Models were applied to six tropical tree species with differing biological attributes: *Oxandra asbeckii*, *Eperua falcata*, *Eperua grandiflora*, *Dicorynia guianensis*, *Qualea rosea*, and *Tachigali melinonii*. We built species-specific models using a simple method of variable selection based on a latent binary indicator.

Our spatial models showed a close correlation between observed and estimated densities with site spatial structure being correctly reproduced. By contrast, the non-spatial models showed poor fits. Variable selection highlighted species-specific requirements and susceptibility to local conditions. Model comparison overall showed that the SGLM was the most accurate explanatory and predictive model. Surprisingly, zero-inflated models performed less well.

Although the SZIP model was relevant with respect to data distribution, and more flexible with respect to response curves, its model complexity caused marked variability in parameter estimates. In the SGLM, the spatial process alone accounted for zero-inflation in the data. A refinement of the hypotheses employed at the process level could compensate for distribution flaws at the data level. This study emphasized the importance of the HBM framework in improving the modelling of density–environment relationships.

1. Introduction

The population dynamics of tropical tree species involves multiple and heterogeneous processes. These biotic and abiotic processes, such as competition and disturbance, are of a particular impact on the spatial patterns of early life-stages. These patterns integrate not only species preferences, but also some dispersal signal which blurs as mortality filters come into operation (Wang and Smith, 2002). Because of this complexity, and particularly in early life-stages, spatial patterns constitute the subject of studies used to draw ecological inference (Austin, 2002). The analysis of these spatial patterns can be valuably conducted in a modelling approach (Guisan and Zimmermann, 2000).

The modelling approach developed here follows the general framework proposed by Austin (2002) which integrates three interacting conceptual components. First, the *ecological* model addresses ecological theory in a given system. When species distributions are concerned, the individualistic community scheme sets a relevant model in which each species interacts with its environment through intrinsic rules (Guisan and Zimmermann, 2000). Second, the *data* model describes the studied system through designed response and explicative variables. In most studies of tree species distribution, the response variable is presence/absence. Fewer studies tackle the local density of conspecifics, especially in tropical rainforests (but see Svenning et al., 2006). This kind of response variable induces zero-inflation which occurs when the frequency

* Corresponding author.
E-mail address: olivierflores@free.fr (O. Flores).

0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.01.030

of zero observations exceeds that expected in a classical distribution. Also, in tropical forests, marked heterogeneity in space and time makes it difficult to define and measure relevant explicative variables. Indirect explicative variables often serve as proxies that quantify ecological processes and direct (physiological) or resource gradients (Guisan and Zimmermann, 2000). The third and final component, the *statistical* model, defines the relationships between data model variables and the methods used for their analysis.

In this contribution we focused on the *statistical* model of Austin's framework in order to develop models of sapling density that include the issues raised by the *data model*: zero-inflated count data, numerous explicative variables, and spatial autocorrelation. Zero-inflation is a common feature of data in many domains and has recently received particular attention (Martin et al., 2005) in ecology. Null observations have different causes: (i) "structural" zeros relate to the absence of a species in unsuitable habitats or because it is scarce (Welsh et al., 1996), whereas (ii) "random" zeros arise by chance from ecological processes (e.g. dispersal limitation), or sampling or observer error (Martin et al., 2005). True zeros (structural or random) arise from ecological processes, whereas false zeros stem from sampling. True zeros are particularly likely to arise in tropical forests, due to vegetation features: extreme species richness implies low specific densities, even in abundant species, and a high frequency of rare species. Focusing on a particular life-stage may also induce zero-inflation because of low abundance.

Zero-inflated (ZI) models are a special case of finite-mixture models that mix two distributions to account for dispersion in data. ZI models offer statistical robustness and flexibility in the shape of response curves (Flores et al., 2006), a central issue in modelling studies (Guisan and Zimmermann, 2000; Oksanen and Minchin, 2002; Austin, 2007). However, they come at the cost of additional complexity over Poisson models. In the conditional ZI (Hurdle) model, structural and random zeros are modeled together as derived from a binomial process (Ridout et al., 1998). Non-zero data are modeled separately through a truncated Poisson (or negative binomial) distribution (Welsh et al., 1996; Barry and Welsh, 2002; Kuhnert et al., 2005). In the mixture ZI model, structural and random zeros are considered separately (Martin et al., 2005; Flores et al., 2006) in a two-stage process. A binary (Bernoulli) process first determines whether, in a second stage, an observation proceeds from a degenerated null process (leading to structural zeros) or from a Poisson process (possibly leading to random zeros). Finite-mixture models generalize parametric methods in allowing specification of non-classical data distributions (Richardson and Green, 1997). However, various alternative parametric and non-parametric methods are also available for empirical modellers to tackle these statistical issues. Recently applied methods include generalized linear models (GLM, Guisan et al., 2002; Miller and Franklin, 2002; Stephenson et al., 2006), generalized additive models (GAM, Barry and Welsh, 2002; Guisan et al., 2002; Moisen and Frescino, 2002), classification and regression trees (CART, Moisen and Frescino, 2002; Miller and Franklin, 2002), multivariate adaptive regression splines (MARS, Moisen and Frescino, 2002) and artificial neural networks (ANN, Moisen and Frescino, 2002).

Spatial autocorrelation has for many years been recognized as ubiquitous in ecological field data (Legendre, 1993). It challenges the classical statistical hypothesis of observations being independent. At the same time, explicit modelling of autocorrelation may provide insight into unobserved processes at various scales (Svenning et al., 2006; Miller et al., 2007). It is in tropical forests that local density is most likely to be autocorrelated. Tree species often display clumped spatial patterns at a local scale (Condit et al., 2000), because of limited dispersal, facilitation by conspecifics or patchy habitat requirement. It is noteworthy that such clumping may also induce zero-inflation, for instance in a regular sampling design. Autocorrelation can also be handled in various ways. At a local scale, a random variable often accounts for some dependence between neighboring observations (Lichstein et al., 2002; Miller et al., 2007; Svenning et al., 2006). Alternatives, for instance auto-regressive (AR) models, and particularly the Conditional Auto-Regressive (CAR) model, can account for spatial dependence arising from ecological processes (Lichstein et al., 2002).

Spatial statistical models can become complex when a mixed data distribution and/or mixed effects are addressed. The Hierarchical Bayesian Modelling (HBM) approach is particularly suited to such cases (Clark, 2005). The main advantage of HBM over other approaches is that it accommodates biological complexity into a series of simple conditional models (Wikle, 2003; Clark, 2005) and provides robust parameter estimates (Angers and Biswas, 2003). The classical hypothesis of independence between observations is replaced by conditional independence, given hypotheses on the structure of data covariance. At the same time, the Bayesian paradigm offers attractive advantages by its ability to integrate prior knowledge into a model, through prior distributions (Banerjee et al., 2003), and to provide a posterior parameters distribution instead of estimated values (Clark, 2005).

When multiple processes are likely to influence the response, the selection of variables becomes paramount. Explicative variables can be selected on a subjective basis or with respect to statistical criteria. Most studies dealing with variable selection use stepwise procedures based on the Akaike Information Criterion (AIC). Parameter estimation then lead to difficulties when variables are numerous, collinear, and when effects are low. Again, an HBM approach may be an effective method for dealing with the selection of variables (Clark, 2005). Several methods have been proposed and may be implemented with varying degrees of difficulty (Dellaportas et al., 2002). Here, we adopt a simple method based on a binary latent indicator. Its estimation provides the posterior probability that an explicative variable improves fitting when included in a model (Dellaportas et al., 2000; Ntzoufras et al., 2000).

This paper describes the building of a density model based on a latent CAR layer that drives a spatially structured behavior to a ZI Poisson data layer. It also compares simple and autocorrelated versions of Poisson and Zero-Inflated Poisson models based on selected explicative variables, and addresses model performance and complexity. The issues of zero-inflation, autocorrelation and variable selection are considered within the HBM framework. The models used are applied to six tropical tree species differing in shade-tolerance and dispersal modes in permanent sample plots (PSP) located in French Guiana. Specific emphasis is placed on investigating the effects of the local environment and intraspecific processes on sapling density.

2. Materials and methods

2.1. Study site and focal species

The study was conducted at the Paracou experimental site (5°18′N, 52°23′W) in a *terra firme* rain forest. The site lies in the coastal part of French Guiana and is subject to an under equatorial climate with a wet season and a dry season. A short drier period interrupts the rainy season from March to April.

The site consists of 300 m × 300 m PSP with a 25 m inner buffer zone. In each central 250 m × 250 m square, all trees ≥ 10 cm diameter at breast height (DBH) were identified and georeferenced. Girth at breast height, tree mortality (standing deaths and treefalls) and recruitment over 10 cm DBH have been monitored annually since 1984. Three treatments were applied over the 1986–1988 period combining selective logging of increasing intensity and additional

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

poison-girdling. The study described here focused on four adjacent PSP (an undisturbed control plot and one treated plot in each treatment) and on the period 1986–2003.

Six focal species were studied: one shade-loving species *Oxandra asbeckii* Pulle, R.E.Fr. (Annonaceae), three shade tolerant to mid-tolerant species (*Eperua falcata* Aublet, Caesalpiniaceae, *Eperua grandiflora* Aublet, Benth., Caesalpiniaceae, *Dicorynia guianensis* Amshoff, Caesalpiniaceae), and two light-demanding species (*Qualea rosea* Aublet, Vochysiaceae, *Tachigali melinonii*, Harms, Caesalpiniaceae). *O. asbeckii* is a bird-dispersed species of the understorey, with maximal height of 15 m. *E. falcata* is self-dispersed and *E. grandiflora* is gravity-dispersed; both species occur in the top canopy at a maximal height of 30–35 m (Sabatier, 1983). *D. guianensis*, *Q. rosea* and *T. melinonii* are wind-dispersed species of the top canopy with emergent trees reaching 40 m. *T. melinonii* is the fastest-growing and most light-demanding of the six species.

2.2. Data model: ecological descriptors

In 2002–2003, all plants in the four plots with 1 cm ≤ DBH ≤ 10 cm were sampled and georeferenced. DBH were recorded in 1-cm classes. Because of large differences in growth potential, tropical trees spend varying periods of time in early life-stages. Here, we allowed the sapling stage to be specifically defined in the data model. The sapling stage was limited by a species-specific upper DBH limit accounting for average growth during the post-logging period. Sapling DBH classes corresponded to 1–2 cm for *O. asbeckii*, 1–3 cm for *E. grandiflora*, 1–4 cm for *E. falcata*, 1–5 cm for *D. guianensis*, 1–6 cm for *Q. rosea* and 1–9 cm for *T. melinonii*. Saplings were counted on an exhaustive and regular basis in 10 m × 10 m cells (625 cells per PSP). The observed sapling density in the cells constituted the studied response variable ($n = 2500$).

Explicative variables constituted of 13 descriptors of ecological conditions that control tropical tree species density (Table 1). These variables were derived either from a Digital Elevation Model (DEM) of the site (elevation and slope), or from census data for trees (≥ 10 cm DBH, stand variables), calculated on 20-m radius plots centered on sampling cells. Two static variables described local forest structure in 2002: total basal area and basal area of pioneer taxa. Five variables characterized stand dynamics during both logging (1986–1988) and the following recovery period (1988–2003): four disturbance variables (Table 1) and a variable quantifying gross change in total basal area over the recovery period. The local disturbance regime was characterized by the mean and standard deviation of treefall age during the recovery period (Table 1).

Finally, two population variables estimated interactions with surrounding conspecific trees (Table 1) to account for intra-population and inter-life-stage autocorrelation. First, the distance from cell center to the nearest adult estimated saplings potential dispersal distance. Second, the basal area of living conspecific trees (≥ 10 cm DBH) on the 20-m radius plots accounted for intraspecific competition. Adults included mature trees, i.e. trees with a DBH greater than a threshold. DBH at maturity was defined with respect to species status and confirmed by literature data when possible: 10 cm for *O. asbeckii*, 25 cm for *D. guianensis*, and 35 cm for *E. falcata*, *E. grandiflora*, *Q. rosea* and *T. melinonii*. Adults included living trees in 2002 and trees either logged during treatment application or that died naturally during the *recovery* period.

2.3. Spatial HBM using latent CAR

The HBM approach accommodates complexity in a high-dimension model through decomposition into a series of simpler conditional hierarchically defined models (Banerjee et al., 2003; Clark, 2005): at a given level, inference conditionally relies on lower-level hypotheses. Three basic levels are mandatory. First, a *data* level specifies the conditional distribution of the data Z given parameters and underlying processes. The hypothesis of conditional independence between observations replaces the classical hypothesis of complete independence. Second, a *process* level specifies the conditional distribution of processes given their own parameters. Third, a *parameter* level specifies the prior distributions of remaining parameters (Wikle, 2003). The purpose of the Bayesian analysis is then to estimate the posterior distribution of the parameters conditional on the data.

A major issue in spatial modelling is to describe correctly the covariance structure of the data. In the sections below, we present a Zero-Inflated Poisson (ZIP) model and its spatial version. The spatial ZIP (SZIP) includes fixed effects and a spatially structured random effect (Fig. 1 a) which models autocorrelation in the response that cannot be explained by fixed effects only. We then briefly describe spatial Poisson models. Finally, we focus on the selection of variables (Fig. 1b), and model calibration and comparison using four criteria.

We modelled the distribution of sapling density as a special case of finite mixture distribution, i.e. the ZIP distribution (Lambert, 1992). In the mixture ZIP model, the distribution of observed data Z follows a mixture of a zero-point mass distribution (modelling structural zeros) and a Poisson distribution $\mathcal{P}(\lambda)$. The model assigns an unknown mass of ω ($0 \leq \omega \leq 1$) to structural zeros and a mass of $(1-\omega)$ to the Poisson distribution. The probability function of the model is

$$\mathbb{P}(Z = z_i|\omega, \lambda) = \begin{cases} \omega + (1-\omega)\mathcal{P}(Z = 0|\lambda) & \text{if } z_i = 0 \\ (1-\omega)\mathcal{P}(Z \neq 0|\lambda) & \text{if } z_i > 0, i = 1, \ldots, n \end{cases}$$

Table 1
Explicative variables derived from a Digitalized Elevation Model (DEM) of Paracou or from census data of trees ≥ 10 cm DBH (*units* in brackets).

Type	Label	Description	Period
Topography	Ele	Elevation (m)	–
	Slo	Slope (°)	
Structure	G_{pio}	Basal area of pioneer taxa (m^2)	2002
	G_{tot}	Total basal area (m^2)	
Logging disturbance	M_{tfL}	Basal area lost in treefalls (m^2)	1986–1988
	M_{sdL}	Basal area lost in standing deaths (m^2)	
Post-logging dynamics	M_{tfR}	Basal area lost in treefalls (m^2)	1989–2002
	A_{tf}	Mean age of treefalls (year)	
	SD_{tfR}	Standard deviation of treefalls ages (year)	
	M_{sdR}	Basal area lost in standing deaths (m^2)	
	dG	Change in basal area (m^2)	
Population variables	dna	Distance to nearest adult (m)	2002
	G_{con}	Basal area of conspecific trees ≥ 10 cm DBH (m^2)	

The period indicates calculus years: 1986–1988 (*logging*) or 1989–2002 (*recovery*). Structure and population variables were calculated in 2002.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

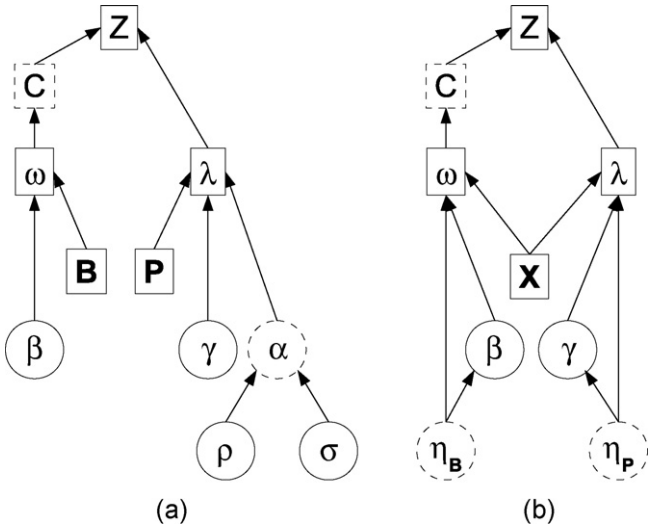


Fig. 1. Directed acyclic graphs of the most complete models: (a) Zero-Inflated Poisson model with random spatial effect, (b) Zero-Inflated Poisson model with binary indicator for variable selection. The models are presented in the four-level HBM scheme including data, process, parameter and hyperparameter levels (Wikle, 2003). Observed or deterministic (defined through an equation, not a distribution) variables are in rectangles. Unknown variables and unknown parameters are in circles. Dashed lines indicate latent variables. Z, observed local sapling density; C, latent binary variable; matrices of explicative variables: $\mathbf{X}$, complete matrices used in variable selection (see Table 1), $\mathbf{X_P}$ and $\mathbf{X_B}$, matrices of selected variables respectively for the Poisson distribution with intensity λ and the binomial distribution with probability ω; γ and β, regression coefficients; $\eta_\mathbf{B}$and $\eta_\mathbf{P}$, latent binary indicators used in variable selection; α, random spatial effect assigned a CAR prior with parameters (ρ, σ; see text for details).

where n is the number of sampling cells, or, using the mixture formulation

$$\mathbb{P}(Z|\omega,\lambda) = \omega \times \delta_0(Z) + (1-\omega)\mathcal{P}(Z|\lambda)$$

where $\delta_0(Z)$ is the Dirac distribution at zero. Here, we introduce a latent (unobserved) random binary variable, C, indicating whether the response Z is structurally null or not. C is modelled as the outcome of a Bernoulli process: $C = 1$ leads to structural zeros (i.e. structural absence), and $C = 0$ indicates that Z follows a Poisson distribution. From Bayes' theorem, the mixture distribution can be expressed as the joint distribution of (Z, C):

$$\mathbb{P}(Z, C|\omega,\lambda) = \mathbb{P}(Z|C=\mathbf{c},\omega,\lambda)\mathbb{P}(C=\mathbf{c}|\omega) = \omega^{\mathbf{c}}[(1-\omega)\mathcal{P}(Z|\lambda)]^{1-\mathbf{c}}$$

At the process level, ω, the probability of a zero being structural, and λ, the intensity of the Poisson process, depends on fixed effects measured by explicative variables through canonical link functions (McCullagh and Nelder, 1989):

$$\text{logit}(\omega) = \mathbf{B}\gamma + \mu \quad (1)$$

$$\log(\lambda) = \mathbf{P}\beta + \alpha \quad (2)$$

where μ and α are two intercepts, $\mathbf{B}$ and $\mathbf{P}$ are two matrices of selected explicative variables (with variables in common or not), and γ and β are two unknown vectors of regression parameters.

In our HBM approach, we extended this ZIP formulation to account for autocorrelation between neighboring observations. We considered that the response variable Z is spatialized, and measured at locations $\mathbf{s}$: $Z = Z(\mathbf{s})$. We assumed that $\alpha(\mathbf{s})$ is a random spatial effect resulting from a spatially structured but unobserved process (see Wikle, 2003). In the SZIP model, at the process level, the Poisson process intensity $\lambda(s)$ thus depended on fixed effects and a random spatial effect:

$$\log[\lambda(\mathbf{s})|\beta,\alpha(\mathbf{s})] = \mathbf{P}\beta + \alpha(\mathbf{s})$$

The intensity of the spatial process, $\alpha(\mathbf{s})$, can be viewed as the spatial component of λ when fixed environmental effects are taken apart: $\alpha = \log(E[y]) - \mathbf{P}\beta$. It is modeled here as a Gaussian random field over a lattice. We used a CAR model (Besag, 1974) for $\alpha(\mathbf{s})$ because observations were sampled on a regular grid and we wanted to account for local autocorrelation. Given a focal location and its neighborhood, the CAR model is interpreted as follows: if the response in the neighborhood gives higher than expected values based on explicative variables, then the focal response will also be locally higher than the expected value. $\alpha(\mathbf{s})$ followed a Gaussian distribution given intensities in a neighborhood:

$$\alpha(s_i)|\alpha(s_j)_{j\in v_i} \sim \mathcal{N}\left(\rho\sum_{j\in v_i} w_{ij}\alpha(s_j), \sigma^2\right), \quad i = 1,\ldots,n \quad (3)$$

where ρ and τ are two unknown parameters, and (w_{ij}) is a set of spatial weights defining neighborhood relationships (see Banerjee et al., 2003; Wall, 2004 for definition). ρ is a spatial dependence parameter measuring the strength of the relationship between the value of α in a focal cell s_i and in its neighborhood v_i. σ^2 is the conditional variance. For each cell, we used a Moore neighborhood (the chess king's move).

Finally, the HBM structure of the SZIP model is (Fig. 1a)

data level :	$Z(s)\|\lambda(s) \sim \mathcal{ZIP}[\lambda(s), \omega(s)]$
process level :	$\text{logit}[\omega(\mathbf{s})\|\mu,\gamma] = \mathbf{B}\gamma + \mu$
	$\log[\lambda(\mathbf{s})\|\beta,\alpha(\mathbf{s})] = \mathbf{P}\beta + \alpha(\mathbf{s})$
parameter level :	priors for γ, β and α,
hyperparameter level :	priors for ρ and σ

It is straightforward to obtain spatial and non-spatial Poisson models from this structure.

2.4. Selection of variables

Our selection method, based on that presented in Dellaportas et al. (2002) and Ntzoufras et al. (2000), uses a binary latent variable that indicates which explicative variables are included or not in the model. Let η be the binary latent variable of length p, the number of candidate variables ($p = 13$, Table 1), so that $\eta_j = 1$ indicates that the jth variable is included in the model ($j \in 1,\ldots,p$), whereas $\eta_j = 0$ excludes the variable. A given model is thus characterized by an associated vector η, an additional parameter. The linear predictor $\mathbf{B}\gamma$ in Eq. (1) becomes

$$\sum_{j=1}^{p} X_{ij}\gamma_j\eta_{\mathbf{B}j}, \quad i \in 1,\ldots,n$$

or in matrix form $\mathbf{X}(\gamma \cdot \eta_\mathbf{B})$ where $\cdot$ indicates the dot product, $\mathbf{X}$ is the complete matrix of explicative variables of dimensions (n, p) and the subscript $\mathbf{B}$ refers to the binomial distribution (Fig. 1b). A similar modification applies to the linear predictor $\mathbf{P}\beta$ in Eq. (2). A major benefit of this approach is that the variables space dimension remains constant during the selection unlike in Reversible Jump approaches (Richardson and Green, 1997).

In theory, it is possible to select fixed effects in models with spatial autocorrelation. In practice, several difficulties are encountered. First, parameter inference requires to develop a complex algorithm whose convergence can be difficult to assess. Second, the inclusion of a spatial effect raises identifiability issues: the random effect could counterbalance fixed effects. Third, the selected variables, together with the associated fixed effects, are generally different in spatial models and their non-spatial counterpart (Kneib et al., 2008). In this contribution, we compared fixed effects with and without a random spatial effect, which requires explicative vari-

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

ables to be the same in both cases. For these reasons, our variable selection was performed without spatial effect (see Fig. 1b).

2.5. Prior choice

Let $\theta = (\eta, \beta, \gamma, \mathbf{c}, \alpha, \rho, \sigma)$, the complete set of unknown parameters and latent variables in the most complete model. At the parameter level, the definition of weakly informative priors for θ components finalizes the definition of the different models.

- For η, we retained a p-binomial distribution

$$\pi(\eta) = \prod_{j=1}^{p} \tau_j^{\eta_j}(1-\tau_j)^{1-\eta_j}$$

  where τ_j is the probability that the jth variable is present in the model. When no a priori information is available, $\tau_j = (1/2), \forall j \in \{1, \ldots, p\}$, and then $\pi(\eta) = 2^{-p}$.
- With regard to regression parameters (γ, β), two cases were possible. In the selection case, we considered a partition of γ for instance into $(\gamma_\eta, \gamma_{\setminus\eta})$, where γ_η and $\gamma_{\setminus\eta}$ correspond to variables that respectively are included in and excluded from the model. The prior of $\gamma|\eta$ was partitioned into a model prior $\pi(\gamma_\eta|\eta)$ and pseudoprior $\pi(\gamma_{\setminus\eta}|\eta)$ (see Dellaportas et al., 2002). A symmetrical definition followed for β. Without selection, Gaussian priors $\mathcal{N}(0, 100)$ were assumed for γ and β.
- The spatial random effect, α, was assigned a CAR prior as defined previously. The prior for the spatial association coefficient, ρ, was uniform. To ensure that the CAR model has a proper distribution, the ρ parameter needs to be constrained to the interval $[1/\lambda_{min}, 1/\lambda_{max}]$ where λ_{min} and λ_{max} are the minimal and maximal eigenvalues of $\mathbf{D}_w^{-(1/2)}\mathbf{W}\mathbf{D}_w^{-(1/2)}$ (see Banerjee et al., 2003 for details). For $1/\sigma^2$, we used a weakly informative Inverse Gamma distribution $\mathcal{IG}(0.1, 0.1)$.
- In our ZIP models, we used a n-binomial distribution for the latent class variable, C.

2.6. Model estimation and comparison

Four models were retained for each species: a simple GLM, a Spatial Generalized Linear Mixed (SGLM) model, a non-spatial ZIP model, and a SZIP model. We inferred the posterior distribution of θ, $\pi(\theta|\mathbf{z})$ using a Monte-Carlo Markov Chain (MCMC) algorithm. Simulations consisted in sampling θ components along a Markov Chain through a hybrid sampling algorithm of Metropolis-Hastings-within-Gibbs steps (see Agarwal et al., 2002 for a parallel approach).

Model fitting in each species–model combination consisted in two stages. Explicative variables were first selected for each species separately without a spatial effect. We retained variables for which the posterior mean of the corresponding components in $\hat{\eta}$ was greater than 0.75. This indicated that these variables had been retained at least three times out of four along the Markov Chain. In a second stage, regression and spatial parameters were estimated in another MCMC run that included only the selected variables. Each stage consisted of 250,000 iterations from which we discarded a 50,000-iterations burn-in sample. Routines were implemented in C language and run under R (R Development Core Team, 2008). Other analyses were also performed with R.

Predictive power was assessed by simulating independent datasets, which bypasses the need for calibrative and predictive datasets. In HBM, a common problem with model comparisons is the number of degrees of freedom (or effective parameters). Spiegelhalter et al. (2002) suggested comparing hierarchical models by means of a Deviance Information Criterion (DIC) based on deviance moments. However, DIC is not invariant to model parameterization (Spiegelhalter et al., 2002; Celeux et al., 2006; Raftery et al., 2007). In this work, we used the classical Spearman's correlation coefficient, and three Bayesian comparison approaches that are independent of model parameterization (see Appendix B for details about criteria).

The first Bayesian criterion, AICM, is an extension of AIC to Monte-Carlo inference based on the Bayes Factor (Raftery et al., 2007). Second, we computed the posterior predictive loss described by Gelfand and Ghosh (1998), D_1, using replicate data conditional on the posterior distribution of observations (see Appendix B for a detailed definition of criterion). Third, we calculated a posterior predictive p-value (p_{ppc}, see Appendix B) based on the posterior predictive check described by Gelman et al. (1996), which also requires simulated replicates of the data. Values of p_{ppc} that are close to 0 or 1 tend to indicate model rejection (Gelman et al., 1996). The best model should give a p_{ppc} of 0.5. Spearman's coefficient and AICM reflect model goodness-of-fit, whereas D_1 and p_{ppc} address model predictive power (Guisan and Zimmermann, 2000; Banerjee et al., 2003).

3. Results

3.1. Observed densities

Zero-inflation varied across species, with zero-frequencies between 58% for *O. asbeckii* and 87% for *T. melinonii*, compared with 40.3% and 78.2% expected for a Poisson distribution with intensity equal to the average observed density (Table 2). *O. asbeckii* was the most abundant species with 2271 identified saplings. By contrast, *D. guianensis* and *T. melinonii* were the lowest in total numbers (615 and 616) and showed the lowest maximal densities (8 and 11). *Q. rosea* was the most abundant species locally (max.: 34, tot.: 1197) and also the most variable in density. *E. falcata* and *E. grandiflora* occurred in 17% and 20% of the cells, respectively, with 17 and 11 saplings at maximal densities (tot.: 807 and 861).

3.2. Model comparison

Regarding the models' explicative power, the spatial models (SGLM, SZIP) showed a closer agreement between observations and fitted densities that did the non-spatial models (ZIP and GLM) with respect to Spearman's correlation coefficient (ρ_s, Table 3). The spa-

Table 2
Outline of sapling density data for the six focal species at the site.

	O. asbeckii	*E. falcata*	*E. grandiflora*	*D. guianensis*	*Q. rosea*	*T. melinonii*
Σ	2271	807	861	615	1197	616
Max	15	17	11	8	34	11
λ_{obs}	0.908	0.323	0.344	0.246	0.479	0.246
V_{obs}	1.570	1.005	0.923	0.735	1.915	0.870
f_0	58.3	83.3	80.4	84.5	84.6	86.9
$\mathcal{P}_0$	40.3	72.4	70.9	78.2	61.9	78.2

Σ: total number of saplings, Max: maximal observed sapling density in a 10 m × 10 m cell, λ_{obs}, V_{obs}: observed mean and variance of sapling density, f_0: observed frequency of zero counts in sapling density, $\mathcal{P}_0$: expected frequency of zero counts in a Poisson distribution with intensity λ_{obs}.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

Table 3
Comparison statistics of estimated models.

	α	ρ_s		|AICM|		D_1		p_{ppc}	
		P	ZIP	P	ZIP	P	ZIP	P	ZIP
O. asbeckii	∅	0.43	0.44	6556	6135	4756	6000	$\simeq 0$	$\simeq 0$
	CAR	**0.76**	0.74	**6028**	6077	**3379**	3696	**0.73**	0.76
E. falcata	∅	0.48	0.49	3072	2742	1896	2249	$\simeq 0$	$\simeq 0$
	CAR	**0.62**	0.60	2776	**2699**	**1203**	1396	**0.59**	0.74
E. grandiflora	∅	0.41	0.42	3403	**3184**	1743	2097	$\simeq 0$	$\simeq 0$
	CAR	**0.64**	**0.63**	3248	3309	**1274**	1329	**0.67**	**0.66**
D. guianensis	∅	0.28	0.29	3187	2797	1267	1691	$\simeq 0$	$< 10^{-1}$
	CAR	**0.60**	0.57	**2672**	3073	**905**	1153	**0.75**	**0.74**
Q. rosea	∅	0.32	0.32	4523	3390	4675	8757	$\simeq 0$	$\simeq 0$
	CAR	**0.69**	**0.69**	**2591**	2626	**1768**	1858	**0.64**	0.68
T. melinonii	∅	0.20	0.19	3408	2820	1546	2306	$\simeq 0$	$\simeq 0$
	CAR	**0.55**	**0.55**	**2498**	2531	**918**	**919**	**0.60**	**0.61**

ρ_s: Spearman correlation coefficient between observations and fitted values, |AICM|: absolute value of the Akaike Information Criterion Monte-Carlo (Raftery et al., 2007); all computed values were negative, D_1: variance-orientated value of the posterior predictive loss (Appendix B with $k = 1$, Gelman et al., 2004). p_{ppc}: posterior predictive p-value; the closer to 0.5, the better the predictive power of the model (Gelman et al., 1996). The α column differentiates results from non-spatial (∅) and spatial models with a CAR-prior random effect. P indicates Poisson-distributed models, and ZIP indicates Zero-Inflated Poisson models. For each species and each statistics, bold numbers indicate the best models. $\simeq 0$ indicates values $< 10^{-4}$.

tial models gave lower absolute AICM values than the non-spatial models, except in *E. grandiflora*. In the non-spatial models, the marked variability of the log-likelihood along the Markov Chain induced high values for |AICM|. Absolute values of AICM were lowest for SGLM in four species, for the SZIP model in *E. falcata*, and for the ZIP in *E. grandiflora* (Table 3). All ZIP models performed better than the Poisson models (GLM) in the non-spatial case, whereas for spatial models, SZIP gave higher values than SGLM, except in *E. falcata* (Table 3).

Regarding the models' predictive power, the spatial models also performed better than their non-spatial counterpart with regard to the posterior predictive loss function, D_1, and the posterior predictive check, p_{ppc} (Table 3). The Poisson models performed better than the Zero-inflated models, but in all species, the non-spatial models gave values of p_{ppc} that were close to 0, indicating model rejection. Overall, SGLM performed best across all models, and this with respect to both D_1 and p_{ppc} (Table 3).

3.3. Comparison of fitted vs. observed patterns

Moran's I (I_M) was calculated as an indicator of local dependence in sapling density. Here, we used the same neighborhood definition as in the CAR model. All observed spatial patterns showed positive I_M values (Fig. 2) with low variance ($< 10^{-3}$, not shown) indicating positive autocorrelation. Overall, I_M values were higher for fitted than for observed patterns. This finding shows that the modelling process tended to smooth fitted distributions. Still, I_M values in the spatial models (SGLM and SZIP) were closer to observed values than to GLM and ZIP values (Fig. 2). In the non-spatial case, the models also failed to account for local maxima in sapling density (Appendix C).

Empirical variograms were calculated in order to analyze spatial patterns at the site scale. Major change in variograms slopes indicated clumps at various scales (Fig. 3, solid lines). A steep increase was observed up to about 50 m for *D. guianensis* and *E. grandiflora*, and up to about 100 m for *O. asbeckii*. Variograms for *Q. rosea* and *E. falcata* showed a slow increase up to 200 m, with higher variability for *Q. rosea*. The variograms for both species increased after 400 m due to isolated clumps (see maps in Appendix C). In *T. melinonii*, the variogram showed a steep increased in the first 30 m, but the overall spatial structure was less marked in this species.

Variograms calculated on model residuals showed how the models accounted for the spatial structure of sapling density. They were all close to zero and flat in spatial models, indicating no residual autocorrelation (Fig. 3). Overall, the spatial models were able to reproduce the spatial structure of sapling density at the local scale. The spatial structure was also well reproduced at the site scale (see maps in Appendix C). With regard to the non-spatial models (GLM and ZIP), the poor agreement between observed and fitted values induced highly autocorrelated residuals.

3.4. Variables selected and effects

The number of explicative variables chosen during the selection phase ranged from 5 in *D. guianensis* (Fig. 4) to 11 (Poisson models in *O. asbeckii*). Most of the variables selected were common across Poisson and zero-inflated models although differences arose in all species (Fig. 4). Overall, the selection procedure retained fewer explicative variables in zero-inflated than in Poisson models. Some variables retained in Poisson models had no influence on sapling

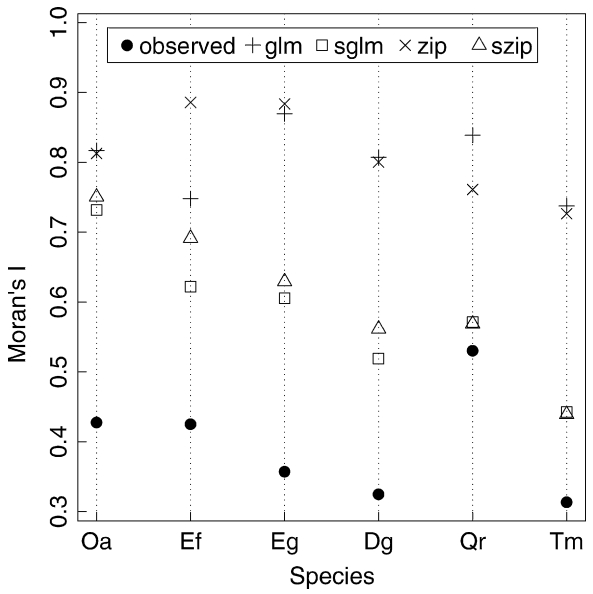


Fig. 2. Moran's I of observed and fitted sapling patterns.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

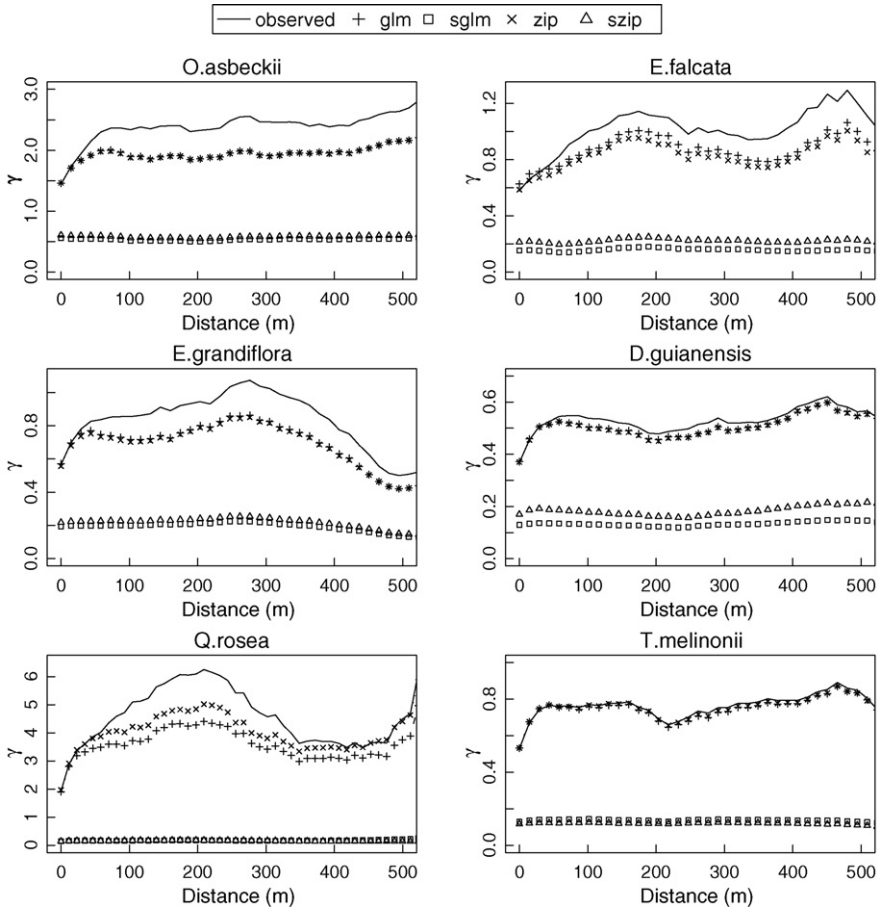


Fig. 3. Spatial structure at the site scale. For each species, the solid line shows the empirical variogram of sapling density (observed), while symbols show variograms calculated on the residuals of the four models (GLM, SGLM, ZIP, SZIP).

density when the zero-inflated distribution was used (e.g. G_{pio} in *E. falcata*, Fig. 4). In fewer cases, variables not retained in the Poisson models were retained in the zero-inflated models (e.g. G_{con} in *E. grandiflora* and *D. guianensis*, G_{con} in *D. guianensis*, G_{pio} in *Q. rosea* and SD_{tfR} in *T. melinonii*).

Each of the 13 explicative variables was retained at least once during the selection phase, and thus partly explained sapling density. Topographic variables, elevation, slope or both, were retained in *O. asbeckii*, *E. grandiflora*, *D. guianensis* and *Q. rosea* (Fig. 4). Structural variables (G_{con} and G_{pio}) were retained in all models except zero-inflated models in *E. grandiflora* and Poisson models in *Q. rosea*. At least one variable characterizing disturbance was selected in all models. Population variables were also retained in all models except in the SZIP in *T. melinonii* (Fig. 4). dna was not retained only in *T. melinonii*.

The comparison of spatial and non-spatial models with similar distributions showed that parameter estimation was substantially altered when autocorrelation was included. The effects of variables, measured here by the posterior mean of the associated regression parameters, generally decreased or reached zero (Fig. 4). Here, we focus on *D. guianensis* which had the fewest selected variables. In the best model for this species (SGLM), the most influent variables were elevation (*Ele*) and distance to nearest adult (*dna*). These had respectively a positive and a negative (decrease with increasing distance) influence on sapling density. These findings indicate a preference for an upper-slope/plateau position and limited dispersal around adults. The other variables retained were SD_{tfR}, G_{tot}, and G_{pio}. The sign of effects indicated that sapling density was more elevated in cells where treefalls were scattered over time, with a low total basal area and a low basal area of pioneer taxa, suggesting conditions of intermediate disturbance intensity.

4. Discussion

This study compared spatial Zero-Inflated Poisson models of sapling local density in a tropical forest with classical GLMs. Overall, model performance was enhanced when accounting for spatial dependence. The CAR model proved well-suited to account for autocorrelation between adjacent cells. The conditional nature of the CAR model makes it relevant for HBM, and HBM does appear to be particularly well adapted in our context. In the spatial models, the residuals appeared to be uncorrelated, showing that the spatial structure of sapling density was relevantly addressed at the local scale. Posterior estimates of the dependence parameter (ρ) were close to its space boundary, suggesting that alternative models could be used. For instance, the Simultaneous Auto-Regressive (SAR) model is formally equivalent to a CAR, but with a different covariance structure (see Keitt et al., 2002; Wall, 2004 for comparisons). Other possibilities include geostatistical models which primarily rely on a continuous description of space. However, auto-regressive models are better suited for the study of area-based data, especially on a regular lattice (Banerjee et al., 2003).

The SGLM showed greater explicative and predictive power than the other models in our case study. Adding a latent spatial effect at the process level of HBM was sufficient to handle both spatial autocorrelation and zero-inflation. We assumed that this was possible because the zero observations were autocorrelated in space (see maps in Appendix C). We suspect that a zero-inflated model would

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

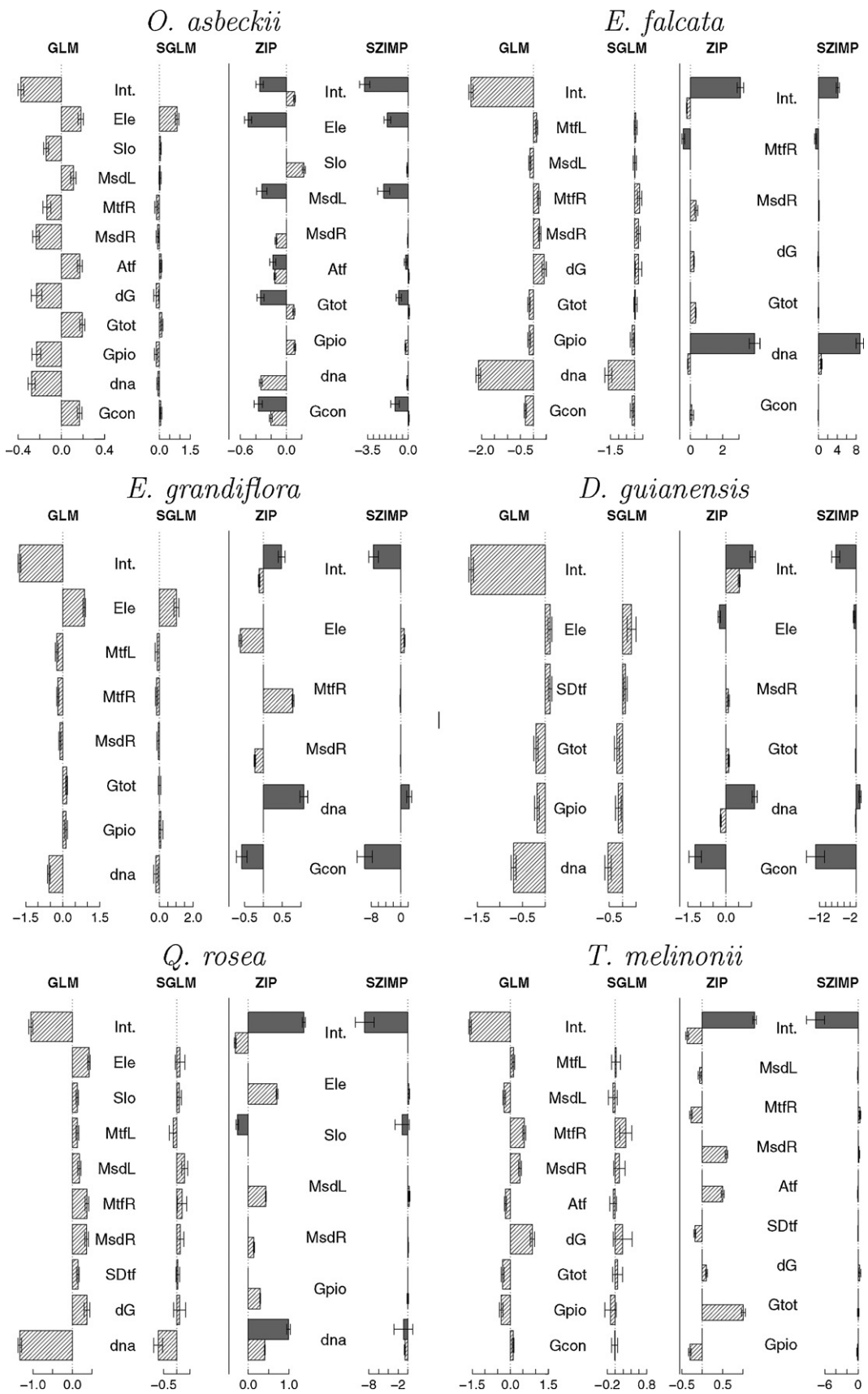


Fig. 4. Explicative variables: selection and effects. The figure shows, for each species and each of the four studied models, the posterior means and standard deviation intervals of regression parameters associated with selected variables (see text for the variable selection procedure and Table 1 for labels, Int.: intercepts). Shaded bars relate to variables included in the Poisson distribution (matrix **X**), filled bars relate to coefficients of variables included in the binomial distribution (matrix **B**).

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

be more efficient in cases of data with uncorrelated structural zeros.

In the SZIP model, we included autocorrelation in the Poisson process, which produces random zeros and non-zero counts. An alternative approach could account for dependence in the probability of observations to be structural zeros (ω). This may be justified in species with strong habitat specificity, for instance in species exclusively found in waterlogged areas. However, such model structure leads to instability and poor parameter estimates (Agarwal et al., 2002). Specific sophisticated algorithms are required to address this issue.

Zero-inflation may be induced by a number of causes in vegetation data. These include scarcity of the studied plants and sampling variability, but also ecological constraints such as habitat unsuitability or marked clumping. ZIP models have the advantage at accounting for these processes and they also allow flexibility in the shape of the response curve, a critical issue in studies of species patterns (Guisan and Zimmermann, 2000; Oksanen and Minchin, 2002). The two-component or Hurdle model is often advocated when facing the mixture specification because parameter interpretation is easier in this case. We preferred the mixture specification for three reasons. First, in the mixture case, the response curve to a given predictor can be easily calculated (Flores et al., 2006). Second, assuming that the processes leading to zero and non-zero data are independent may not be relevant to the ecological model. For instance, habitat suitability is not a binary factor: individuals surviving in transient habitats imply non-null density. Likewise, dispersal may induce structural zeros beyond a limiting distance, though infrequent long distance dispersal events occur. Dispersal also implies random zeros as seeds do not saturate a tree's influence area, because of stochasticity. Third, the mixture specification separates effects leading to structural and random zeros. Selected variables can influence either the binary, or the Poisson process, or both.

The saplings in our study appeared to be clumped, which may be due to limited dispersal around adults (Svenning, 2001), clumped seed dispersal (Howe, 1989; Russo and Augspurger, 2004), or a survival response to patchy resources (Dalling et al., 1998). Svenning et al. (2006) interpreted the high contribution made by the CAR component to local density as evidence of strong local dispersal. Clearly, aggregate dispersal is likely to induce local autocorrelation in species patterns. However, we would expect the CAR component to contribute differently across species that display different dispersal modes. No such findings were observed. In light-demanding species (*e.g. T. melinonii*), autocorrelation may reflect unobserved environmental heterogeneity induced by unobserved disturbance events (Nicotra et al., 1999).

In our study, we characterized the environment by means of continuous descriptors of ecological processes such as disturbance. This approach addressed a common issue in modelling, i.e. a disequilibrium between observed patterns and current environmental conditions (Guisan and Zimmermann, 2000; Austin, 2002). Overall, non-null effects were detected in each model-species combination, and selected variables changed across species. Designed variables thus all quantified some aspect of environmental heterogeneity or population process that partly explained sapling density and indicated specific processes. The position of adults influenced sapling patterns in five species. Whereas no such effect was seen for the anemochorous and most light-demanding species *T. melinonii*. Despite mortality filters on earlier stages, a dispersal signal persisted in sapling patterns (Clark et al., 1999; Wang and Smith, 2002). The studied species are known to be rather poor dispersers, like the bird-dispersed *O. asbeckii* (Ulft, 2004), the autochorous and barochorous *Eperua* and *D. guianensis* despite wind dispersal. Rodents secondarily dispersing seeds can increase dispersal distances (Forget, 1992).

Variables of past disturbance patterns were particularly informative, and this is consistent with previous studies of the spatial heterogeneity of light in tropical forests (Nicotra et al., 1999). Overall, disturbance effects were consistent with species shade-tolerance. Species recruitment was differentially affected by disturbance-induced opening of the canopy. Regarding topography, *D. guianensis* and *E. grandiflora* are known to mainly settle on the upper part of slopes and *E. falcata* on bottomlands. Here, *E. falcata* was weakly affected by topography. In this species, population variables were sufficiently informative to mask the effects of physical conditions because of the marked clumping of saplings around adults. This finding raises the issue of covariance between explicative variables. Here, we selected explicative variables from candidates based on an efficient and stable selection method. We used a prior that favors models with $p/2$ variables. Other choices are possible that would take account of covariance between variables. The influence of such priors on statistical and ecological inference remains to be tested.

In modelling studies, the trade-off between model complexity and relevance is a well-known issue. In our case, the SZIP model appeared to be conceptually relevant as it could account for two critical features of data, i.e. autocorrelation and zero-inflation. Empirically, the SGLM model appeared to show the best performance. This finding shows that refining processes addressed at the process level of HBM could compensate for statistical dispersion observed in the data. In other words, a priori required complexity at the data level was not necessary when accurate specification occurred at process level. In a predictive context, statistical simplicity may be preferred. Sophisticated models may nevertheless be required to evidence hidden biological processes.

Acknowledgements

We wish to thank Lilian Blanc, Jean-Gaël Jourget, Pascal Pétronelli (CIRAD, Kourou, French Guiana) and the Paracou field workers who participated in collecting the data. We also extend our thanks to Sylvie Gourlet-Fleury and Hélène Dessard for constructive discussions and helpful comments.

Appendix A. Site map

See Fig. 5.

Appendix B. Model comparison

The Bayes Factor (BF) is among the common approaches used for model comparison. It is based on the integrated posterior harmonic mean of the likelihood:

$$\pi(Z) = \int f(Z|\theta)\pi(\theta)\,d\theta.$$

which can be approximated by the harmonic mean of the likelihood along a standard Markov Chain Monte-Carlo run (Raftery et al., 2007). Although $\pi(Z)$ is consistent as the simulation size increases, its precision is not guaranteed. Raftery et al. (2007) proposed the use of a shifted gamma estimator which leads to modified versions of AIC and BIC. We retained the AICM (M for Monte-Carlo) which addresses model explicative power, and is defined as

$$\mathrm{AICM} = 2(\hat{l} - s_l^2)$$

where $\hat{l}$ and s_l^2 are the mean and variance of the log-likelihood along the chain.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

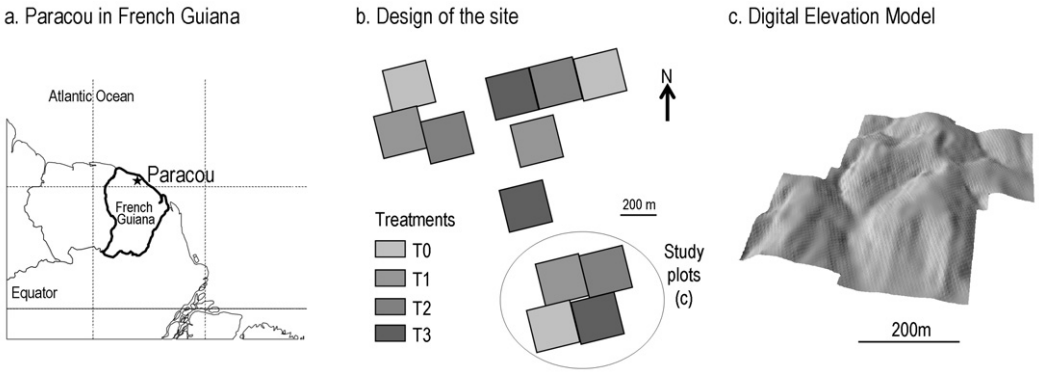


Fig. 5. Location and map of the study site.

An alternative approach is the posterior predictive loss described by Gelfand and Ghosh (1998) which addresses model predictive power. It uses the distribution of replicate data conditional on the posterior distribution of observations (the posterior predictive distribution). We note $\mathbf{z}^{\text{rep}}$ a replicate dataset simulated with sampled values of parameters θ along the Markov chain. These data could have been observed under the studied model with those values of θ (Gelman et al., 1996). Conditional on θ, $\mathbf{z}^{\text{rep}}$ and $\mathbf{z}$ are assumed to be independent. The posterior predictive distribution of replicates is then

$$p(\mathbf{z}^{\text{rep}}|\mathbf{z}) = \int p(\mathbf{z}^{\text{rep}}|\theta, \mathbf{z})p(\theta|\mathbf{z})\,d\theta$$

The best model then minimizes the posterior predictive loss defined as

$$D_k = \frac{k}{k+1}G + P$$

where

$$G = \sum_{i=1}^{n}(\hat{\mu}_i - z_i)^2 \quad \text{and} \quad P = \sum_{i=1}^{n}\hat{\sigma}_i^2$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the mean and variance of the posterior predictive distribution. The loss function D_k reflects the classical compromise between bias and variance, depending on the choice of k: G and P are, respectively, the bias (the goodness of fit) and the variance of the prediction. In the paper, we use the variance-orientated loss function $D_1 = (G/2) + P$ ($k = 1$) as a second Bayesian criterion. In order to estimate D_1, we simulate 100 replicated data for each of 1000 values of θ sampled along the chain ($\theta^{(k)}$, $k = 1, \ldots, 1000$).

Finally, we derived a last criterion of model validation using the posterior predictive check approach (Gelman et al., 1996). This approach also requires simulated replicates of the data. A discrepancy measure based on the residual sum of squares, $T(\mathbf{z}, \theta)$, quantified model fit:

$$T(\mathbf{z}, \theta^{(k)}) = \sum_i (z_i - E[z_i|\theta^{(k)}])^2$$

where (k) indicates values sampled along the chain.

The goodness-of-fit of a model is then evaluated by comparing the posterior distribution of $T(\mathbf{z}, \theta^{(k)})$ with the posterior predictive reference distribution $T(\mathbf{z}^{\text{rep}}, \theta^{(k)})$ (Stern and Cressie, 2000). We quantified the closeness of two discrepancy measures based on parameters estimates and either the observations or a simulated replicate dataset. Graphically, scattering away from the 1:1 line in the plot of $T(\mathbf{z}, \theta^{(k)})$ and $T(\mathbf{z}^{\text{rep}}, \theta^{(k)})$ indicates that data generated by the model greatly differ from the observed data, with respect to T. Numerically, this information can be summarized by a posterior predictive p-value:

$$p_{\text{ppc}} = \mathbb{P}[T(\mathbf{z}^{\text{rep}}, \theta) \geq T(\mathbf{z}, \theta)]$$

Values close to 0 or 1 tend to indicate model rejection (Gelman et al., 1996). In order to estimate p_{ppc}, we use the approximation:

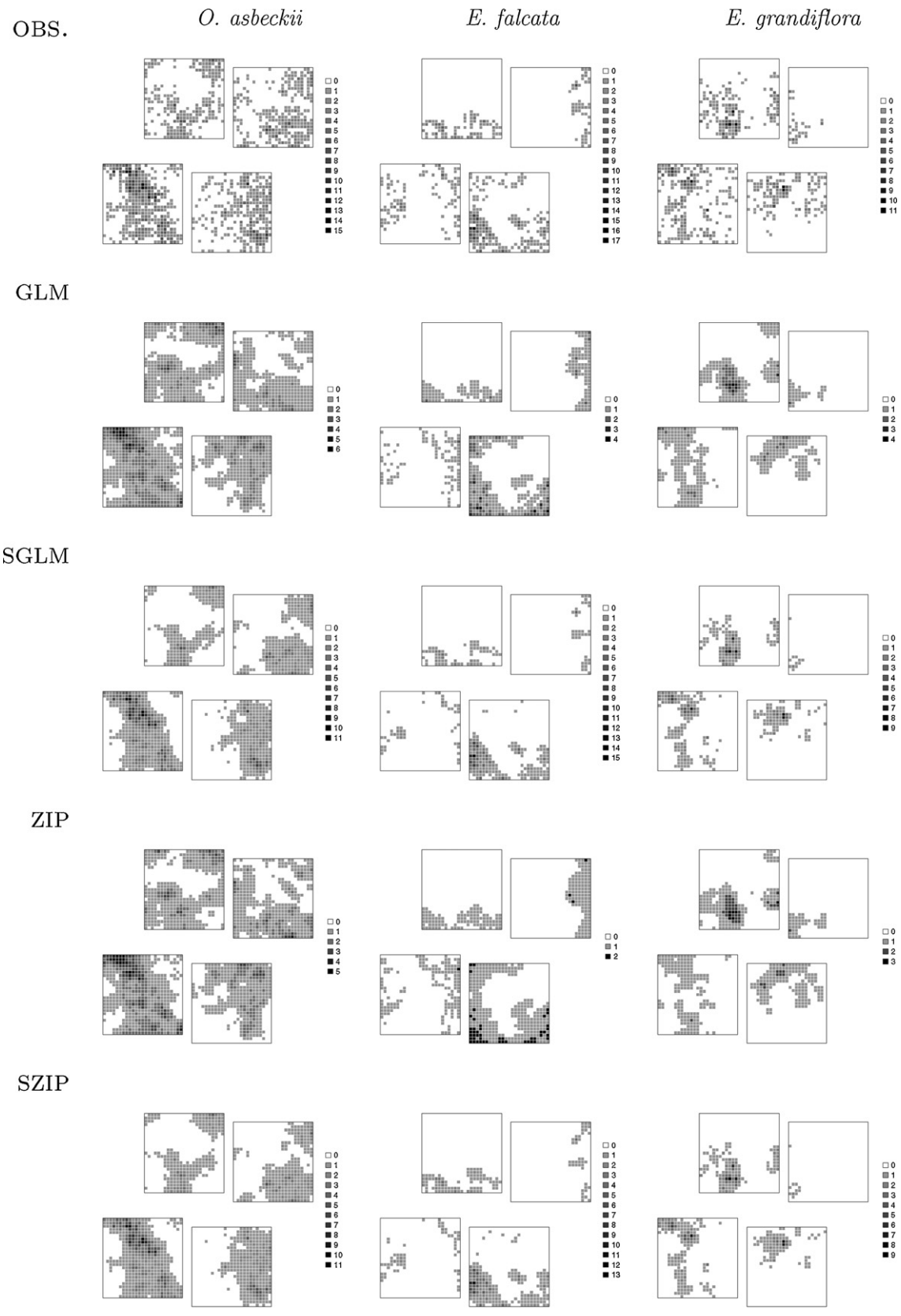
$$p_{\text{ppc}} = \sum_k \sum_j \mathbb{I}_{T(\mathbf{z}_j^{\text{rep}(k)}, \theta^{(k)}) \geq T(\mathbf{z}, \theta^{(k)})}$$

where j indicates a simulated dataset using parameters $\theta^{(k)}$ ($j = 1, \ldots, 100$).

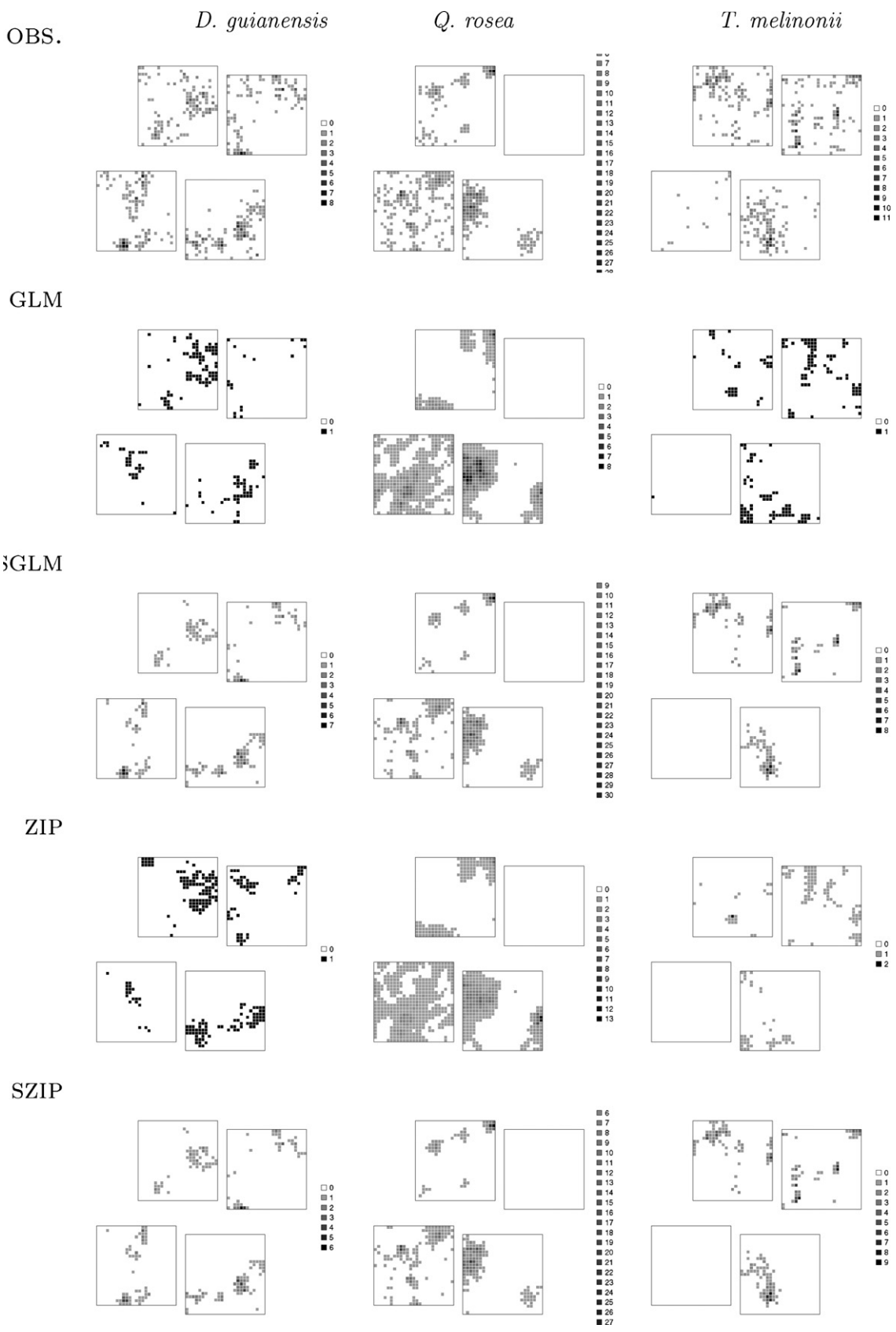
Another common criterion in the comparison of Bayesian models is the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002). However, when hidden structures and random effects are addressed, the definition of the posterior estimates of parameters, $\bar{\theta}$ is not fixed so that DIC depends on model parametrization Spiegelhalter et al. (2002), Celeux et al. (2006), and Raftery et al. (2007), and on a certain focus on the hierarchy (Plummer, 2006). Although Celeux et al. (2006) proposed several versions of the DIC, none seems to be well suited to such cases (see discussion in Celeux et al., 2006 paper).

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

Appendix C. Density maps



Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030



Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

References

Agarwal, D.K., Gelfand, A.E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. Environ. Ecol. Stat. 9 (4), 341–355.
Angers, J.F., Biswas, A., 2003. A Bayesian analysis of zero-inflated generalized Poisson model. Comput. Stat. Data Anal. 42 (1–2), 37–46.
Austin, M., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157 (2–3), 101–118.
Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200 (1–2), 1–19.
Banerjee, S., Carlin, B., Gelfand, A., 2003. Hierarchical modeling and analysis for spatial data. In: Monographs on Statistics and Applied Probability, vol. 101. Chapman & Hall/CRC.
Barry, S., Welsh, A., 2002. Generalized additive modelling and zero inflated count data. Ecol. Model. 157 (2–3), 179–188.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. Ser. B 36, 192–236.
Celeux, G., Forbes, F., Robert, C., Titterington, M., 2006. Deviance information criteria for missing data models. Bay. Anal. 1, 651–674.
Clark, J., 2005. Why environmental scientists are becoming Bayesians? Ecol. Lett. 8, 2–14.
Clark, J., Beckage, B., Camill, P., Cleveland, B., Hillerislambers, J., Lichter, J., McLachlan, J., Mohan, J., Wyckoff, P., 1999. Interpreting recruitment limitation in forests. Am. J. Bot. 86 (1), 1–16.
Condit, R., Ashton, P., Baker, P., Bunyavejchewin, S., Gunatilleke, S., Gunatilleke, N., Hubbell, S., Foster, R., Itoh, A., LaFrankie, J., Lee, H., Losos, E., Manokaran, N., Sukumar, R., Yamakura, T., 2000. Spatial patterns in the distribution of tropical tree species. Science 288, 1414–1418.
Dalling, J., Hubbell, S., Silvera, K., 1998. Seed dispersal, seedling establishment and gap partitioning among pioneer tropical trees. J. Ecol. 86, 674–689.
Dellaportas, P., Forster, J.J., Ntzoufras, I., 2000. Bayesian variable selection using the Gibbs sampler. In: Dey, D.K., Ghosh, S.K., Mallick, B.K. (Eds.), Generalized Linear Models: A Bayesian Perspective. Chemical Rubber Company Press, New York, USA, pp. 273–286.
Dellaportas, P., Forster, J., Ntzoufras, I., 2002. On Bayesian model and variable selection using MCMC. Stat. Comput. 12 (1), 27–36.
Flores, O., Gourlet-Fleury, S., Picard, N., 2006. Local disturbance, forest structure and dispersal effects on sapling distribution of light-demanding and shade-tolerant species in a French Guianan forest. Act. Oec. 29 (2), 141–154.
Forget, P.-M., 1992. Regeneration ecology of *Eperua grandiflora* (Caesalpiniaceae), a large-seeded tree in French Guiana. Biotropica 24 (2a), 146–156.
Gelfand, A.E., Ghosh, S.K., 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85 (1), 1–11.
Gelman, A., Carlin, B., Stern, H., Rubin, D.B., 2004. Bayesian Data Analysis. CRC Press.
Gelman, A., Meng, X.-L., Stern, S., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. 6, 733–807.
Guisan, A., Edwards, T.J., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157 (2–3), 89–100.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2/3), 147–186.
Howe, H., 1989. Scatter- and clump-dispersal and seedling demography: hypothesis and implications. Oecologia 79, 417–426.
Keitt, T., Bornstad, O., Dixon, P., Citron-Pousty, S., 2002. Accounting for spatial pattern when modeling organism–environment interactions. Ecography 25, 616–625.
Kneib, T., Hothorn, T., Tutz, G., 2008. Variable selection and model choice in geoadditive regression models. Biometrics on line.
Kuhnert, P.M., Martin, T.G., Mengersen, K., Possingham, H.P., 2005. Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. Environmetrics 16 (7), 717–747.
Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.
Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecologia 74 (6), 1659–1673.
Lichstein, J., Simons, T., Shriner, S., Franzreb, K., 2002. Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. 72 (3), 445–463.
Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., Possingham, H.P., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol. Lett. 8 (11), 1235–1246.
McCullagh, P., Nelder, J., 1989. Generalized linear models. In: Monographs on Statistics and Applied Probability, vol. 37, Chapman & Hall edition. Chapman & Hall/CRC, London.
Miller, J., Franklin, J., 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecol. Model. 157 (2–3), 227–247.
Miller, J., Franklin, J., Aspinall, R., 2007. Incorporating spatial dependence in predictive vegetation models. Ecol. Model. 202 (3–4), 225–242.
Moisen, G., Frescino, T., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157 (2–3), 209–225.
Nicotra, A.B., Chazdon, R.L., Iriarte, S.V.B., 1999. Spatial heterogeneity of light and woody seedling regeneration in tropical forests. Ecologia 80 (6), 1908–1926.
Ntzoufras, I., Forster, J.J., Dellaportas, P., 2000. Stochastic search variable selection for log-linear models. J. Stat. Comput. Simul. 68 (1), 23–37.
Oksanen, J., Minchin, P., 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecol. Model. 157 (2–3), 119–129.
Plummer, M., 2006. Comment on article by Celeux et al. Bay. Anal. 1 (4), 681–686.
R Development Core Team (2008). R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0.
Raftery, A., Newton, M., Satagopan, J., Krivitsky, P., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Bay. Stat. 8, 1–45.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components. J. Roy. Stat. Soc. Ser. B 59 (4), 731–792.
Ridout, M., Demetrio, C., Hinde, J., 1998. Models for count data with many zeros. In: Proceedings of the International Biometric Conference, Cape Town.
Russo, S., Augspurger, C., 2004. Aggregated seed dispersal by spider monkeys limits recruitment to clumped patterns in *Virola calophylla*. Ecol. Lett. 7 (11), 1058–1067.
Sabatier, D., 1983. Fructification et dissmination en fort guyanaise - L'exemple de quelques espces ligneuses. Doctorat de 3?me cycle, Universit des Sciences et Techniques du Languedoc.
Spiegelhalter, D., Best, N., Carlin, B., Van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). J. Roy. Stat. Soc. Ser. B 6, 583–639.
Stephenson, C., MacKenzie, M., Edwards, C., Travis, J., 2006. Modelling establishment probabilities of an exotic plant, *Rhododendron ponticum*, invading a heterogeneous, woodland landscape using logistic regression with spatial autocorrelation. Ecol. Model. 193 (3–4), 747–758.
Stern, H., Cressie, N., 2000. Posterior predictive model checks for disease mapping models. Stat. Med. 19, 2377–2397.
Svenning, J.-C., 2001. Environmental heterogeneity, recruitment limitation and the mesoscale distribution in a tropical Montane rain forest (maquipucuna, ecuador). J. Trop. Ecol. 17, 97–113.
Svenning, J.-C., Engelbrecht, B.M.J., Kinner, D.A., Kursar, T.A., Stallard, T.A., Wright, S.J., 2006. The relative roles of environment, history and local dispersal in controlling the distributions of common tree and shrub species in a tropical forest landscape, panama. J. Trop. Ecol. 22, 575–586.
Ulft, L.v., 2004. Regeneration in natural and logged tropical rain forest. Modelling seed dispersal and regeneration of tropical trees in Guyana. In: Volume 12 of Tropenbos-Guyana Series. Tropenbos International, Georgetown.
Wall, M., 2004. A close look at the spatial structure implied by the CAR and the SAR models. J. Stat. Plan. Inf. 121, 311–324.
Wang, B., Smith, T., 2002. Closing the seed dispersal loop. Trend Ecol. Evol. 17 (8), 379–385.
Welsh, A., Cunningham, R., Donnelly, C., Lindenmayer, D., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. 88 (1–3), 297–308.
Wikle, C., 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecologia 84, 1382–1394.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

A Hierarchical Bayesian Model for Spatial Prediction of Multivariate Non-Gaussian Random Fields

Biometrics 2010

Biometrics 67, 97–105
March 2011
DOI: 10.1111/j.1541-0420.2010.01415.x

A Hierarchical Bayesian Model for Spatial Prediction of Multivariate Non-Gaussian Random Fields

Pierrette Chagneau,[1,4,*] Frédéric Mortier,[2,] Nicolas Picard,[3,***] and Jean-Noël Bacro[4,****]**

[1]CIRAD, UR Dynamique des forêts naturelles, 34 398 Montpellier, France
[2]CIRAD, UR Diversité génétique et amélioration des espèces forestières, 34 398 Montpellier, France
[3]CIRAD, UR Dynamique des forêts naturelles, Libreville, Gabon
[4]I3M, UMR CNRS 5149, Université de Montpellier 2, 34 095 Montpellier, France
[*]*email:* pierrette.chagneau@cirad.fr
[**]*email:* frederic.mortier@cirad.fr
[***]*email:* nicolas.picard@cirad.fr
[****]*email:* bacro@math.univ-montp2.fr

Summary. As most georeferenced data sets are multivariate and concern variables of different types, spatial mapping methods must be able to deal with such data. The main difficulties are the prediction of non-Gaussian variables and the modeling of the dependence between processes. The aim of this article is to present a new hierarchical Bayesian approach that permits simultaneous modeling of dependent Gaussian, count, and ordinal spatial fields. This approach is based on spatial generalized linear mixed models. We use a moving average approach to model the spatial dependence between the processes. The method is first validated through a simulation study. We show that the multivariate model has better predictive abilities than the univariate one. Then the multivariate spatial hierarchical model is applied to a real data set collected in French Guiana to predict topsoil patterns.

Key words: Count data; Moving average; Ordinal data; Soil; Spatial prediction.

1. Introduction

The prediction of multivariate spatial processes is a major issue in many research areas including biological sciences (McBratney et al., 2000), epidemiology (Golam Kibria et al., 2002), and economics (Chica-Olmo, 2007; Gelfand et al., 2007). In most cases, few data are available as they are expensive to collect. Moreover, data are often of different types. For example, in geological studies, concentrations of elements (continuous variables), granularity (ordinal variables), and coloration (nominal variables) are usually measured to characterize soils. Spatial mapping methods thus have to be able to handle related data of different types. This raises two difficulties: predicting multivariate discrete random fields and modeling the dependence between continuous and discrete spatial processes.

In the univariate case, the prediction of continuous spatial processes has been widely studied and implemented (Cressie, 1991; Wackernagel, 2003). For discrete random fields, methods based on geostatistics and point processes have been developed: disjunctive kriging (Webster and Oliver, 1990), truncated Gaussian random fields (De Oliveira, 2000), object models and Markov random fields (Cressie, 1991; Molchanov, 1997). Recently, new models have been defined particularly to deal with count variables. Wolpert and Ickstadt (1998) proposed modeling count data with a Poisson distribution whose intensity is the unobserved value of a random measure modeled by a gamma process. Diggle, Tawn, and Moyeed (1998) proposed embedding linear kriging methodology in the framework of the generalized linear mixed model where the random effect is modeled by a Gaussian spatial process. This method can predict not only count variables but also Gaussian and ordinal ones. An extension of this methodology, called the geoadditive model, was proposed by Kammann and Wand (2003). This model resulted from the fusion of generalized linear models and additive models (Augustin et al., 2007). Most of these new models are now often described in a hierarchical Bayesian framework (Christensen and Waagepetersen, 2002; Banerjee, Carlin, and Gelfand, 2004). A hierarchical Bayesian approach accommodates complexity in high-dimension models by decomposing a model into a series of simpler conditional levels (Wikle, 2003).

Multivariate spatial processes have been widely studied in recent decades (Cressie, 1991; Wackernagel, 2003). The models proposed are efficient but they require certain restricting assumptions: normality for linear cokriging methods (Cressie, 1991) or isofactorial model assumptions for disjunctive cokriging (Matheron, 1976; Rivoirard, 1991). Modeling the dependence between variables is closely linked to the prediction method chosen. Cokriging methods are based on a full covariance structure model, whereas disjunctive cokriging methods involve hypotheses based on bivariate distributions. In the latter, the determination of the bivariate distributions can be

tedious and the classical isofactorial Gaussian model may be unsuitable. Here we use the alternative procedure proposed by Ver Hoef and Barry (1998).

Many studies have been published on the topic of multivariate spatial models. The intrinsic correlation model is the simplest multivariate covariance model (Matheron, 1965). The coregionalization model generalizes it so that the multivariate correlation structure can be taken into account at different scales of a phenomenon (Matheron, 1965). The latter class of covariance models assumes that the correlation structures for and between each variable are the same up to a constant. Moreover, the choice of each elementary covariance structure in coregionalization models should ensure that the global covariance matrix is positive definite. The use of these approaches is seriously restricted by these two constraints.

Instead, we use the alternative procedure proposed by Ver Hoef and Barry (1998). Barry and Ver Hoef (1996) defined a new family of valid variograms using moving average functions. Ver Hoef and Barry (1998) generalized their approach to the multivariate Gaussian case by convolving white noise processes with moving average functions (Higdon, 2001; Calder and Cressie, 2007). The corresponding covariance matrix has an explicit expression. Moving average constructions are attractive because the variograms obtained are very flexible. The method allows a valid covariance matrix to be obtained even for anisotropic data, as long as the moving average function is well chosen.

Although many studies have been conducted on both problems raised by the prediction of multivariate random fields made up by variables of different types, no method has been proposed to take them into account simultaneously in order to predict such spatial processes. The aim of this article is to propose a new unified approach that can simultaneously model Gaussian, count, and ordinal spatial fields. Our model is based on a hierarchical Bayesian framework. We generalize Diggle et al.'s (1998) method to the multivariate case. In particular, our model can take ordinal spatial processes into account through generalization of the multivariate ordinal probit model to the spatial case (Chaubert, Mortier, and Saint-André, 2008). Our modeling introduces spatial Gaussian latent processes to model spatial dependence. In Section 2, we define the spatial hierarchical model. In Section 3, we describe posterior analysis and the Bayesian implementation of our model. After validating the multivariate model through a simulation study in Section 4, we apply it to predict topsoil patterns from a real data set collected in French Guiana in Section 5. Finally, in Section 6, we draw some conclusions and review the next steps in our research.

2. Spatial Model for Random Variables of Different Types

The spatial model is based on a hierarchical framework like Wolpert and Ickstadt's (1998) model, with three levels. The hierarchical spatial model is specifically designed to take into account variables of different types. The model can be defined for any number K of response variables but, for sake of simplicity and unless otherwise specified, we consider three ($K = 3$) different types of variables (a Gaussian variable, a Poisson variable, and an ordinal variable).

2.1 *First Level of the Hierarchy*

Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ be the n sampled locations. Let $Y_1(\boldsymbol{s}_i)$ be a Gaussian variable at location $\boldsymbol{s}_i$, let $Y_2(\boldsymbol{s}_i)$ be a Poisson variable, and let $Y_3(\boldsymbol{s}_i)$ be an ordinal variable with L categories. In general, if there are separate locations for each variable, one would denote the ith location of the kth variable by $\boldsymbol{s}_{ki}$, but here we restrict ourselves to colocated data. Let $\boldsymbol{Y}_k = (Y_k(\boldsymbol{s}_1), \ldots, Y_k(\boldsymbol{s}_n))', k = 1, 2, 3$ be the vector of the variable Y_k at all sampled locations. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2', \boldsymbol{Y}_3')'$ be the vector of all variables at all sampled locations.

The Gaussian variable $Y_1(\boldsymbol{s})$, the Poisson variable $Y_2(\boldsymbol{s})$, and the ordinal variable $Y_3(\boldsymbol{s})$ depend on centered latent variables $S_1(\boldsymbol{s})$, $S_2(\boldsymbol{s})$, and $S_3(\boldsymbol{s})$, respectively, that are responsible for the spatial dependence. Given $S_1(\boldsymbol{s})$, $S_2(\boldsymbol{s})$, and $S_3(\boldsymbol{s})$, the variables $Y_1(\boldsymbol{s})$, $Y_2(\boldsymbol{s})$, and $Y_3(\boldsymbol{s})$ are conditionally independent. For the Gaussian and Poisson variables, we follow the generalized linear model proposed by Diggle et al. (1998):

$$Y_1(\boldsymbol{s}_i)|\mu_1, S_1(\boldsymbol{s}_i), \nu_1 \sim \mathcal{N}\left(\mu_1 + S_1(\boldsymbol{s}_i), \nu_1^2\right), \tag{1}$$

$$Y_2(\boldsymbol{s}_i)|\mu_2, S_2(\boldsymbol{s}_i) \sim \mathcal{P}\{\exp(\mu_2 + S_2(\boldsymbol{s}_i))\}. \tag{2}$$

where $\mathcal{N}(m, \sigma^2)$ is the Gaussian distribution with mean m and variance σ^2, $\mathcal{P}(\lambda)$ is the Poisson distribution with parameter λ, μ_1 and μ_2 are the overall mean parameters of Y_1 and Y_2, and ν_1^2 is the nugget effect related to the process Y_1. For the ordinal variable, the conditional distribution of $Y_3(\boldsymbol{s})$ is the one that follows from the ordinal probit model. This model reformulates the discrete issue into a continuous problem by introducing a latent Gaussian variable $Z_3(\boldsymbol{s})$ with unit standard deviation (Chib and Greenberg, 1998). A partition of $\mathbb{R}$ in L half-open intervals is defined by a $L-1$-vector of breakpoints, and an equivalence is established between the lth level of $Y_3(\boldsymbol{s}_i)$ and the membership of $Z_3(\boldsymbol{s}_i)$ to the lth interval

$$\begin{aligned}\mathbb{P}(Y_3(\boldsymbol{s}_i) &= l \mid \mu_3, S_3(\boldsymbol{s}_i), \boldsymbol{\alpha}_3) \\ &= \mathbb{P}(Z_3(\boldsymbol{s}_i) \in]\alpha_{3;\,l-1}, \alpha_{3;\,l}] \mid \mu_3, S_3(\boldsymbol{s}_i)),\end{aligned} \tag{3}$$

$$Z_3(\boldsymbol{s}_i) \mid \mu_3, S_3(\boldsymbol{s}_i) \sim \mathcal{N}(\mu_3 + S_3(\boldsymbol{s}_i), 1), \tag{4}$$

where $\boldsymbol{\alpha}_3 = (-\infty, \alpha_{3,1}, \alpha_{3,2}, \ldots, \alpha_{3,L-1}, +\infty)$ is the vector of breakpoints, and μ_3 is the overall mean parameter of Z_3. To ensure that the model is identifiable, we must either assume that $\mu_3 = 0$ (and then $\boldsymbol{\alpha}_3$ is identifiable), or that $\alpha_{3,1} = 0$ (and then μ_3 and the $L-2$ remaining breakpoints are identifiable; Albert and Chib, 1993; Cowles, 1996). From now on, we use the latter parameterization. Expressions (1)–(4) form the first level of the hierarchical model.

This definition for $K = 3$ readily extends to any number of variables. If the random field comprises several Poisson variables, they are all conditionally independent given the latent spatial processes S_k. If the random field comprises several Gaussian variables, a covariance matrix between them (the same at all locations) can be considered. To deal with several ordinal variables, we rely on the multivariate ordinal probit model (Chib and Greenberg, 1998; Chen and Shao, 1999, see also Web Appendix A for a precise definition). The latent variable $Z_3(\boldsymbol{s})$ then is multivariate Gaussian with correlation matrix $\boldsymbol{R}$ (the same at all locations). The matrix $\boldsymbol{R}$ is not a covariance matrix for identifiability reasons (Chib and Greenberg, 1998; De Oliveira, 2000). Although it is possible to

estimate the correlation matrix $\boldsymbol{R}$, we did not focus on this task here and made the simplifying assumption that $\boldsymbol{R}$ is the identity.

2.2 *Second Level of the Hierarchy*

The spatial dependence between the processes Y_k is modeled by the latent Gaussian processes $S_k,\ k = 1, 2, 3$. The processes S_k are built according to the moving average construction proposed by Ver Hoef and Barry (1998), that is to say by convolving a moving average function with a mixture of white noise processes. Let $V_k,\ k = 1, 2, 3$ be a linear combination of white noise processes

$$V_k(\boldsymbol{x} \mid \rho_k) = \sqrt{1-\rho_k^2}W_k(\boldsymbol{x}) + \rho_k W_0(\boldsymbol{x}),$$

where $W_k,\ k = 0, 1, 2, 3$, is a white noise process, and $\rho_k, k = 1, 2, 3$, belongs to the interval $[-1; 1]$. The process W_0 induces dependence between the V_k processes because

$$\begin{aligned}&\mathrm{Cor}\left(\int_{\mathbb{R}^2} V_k(\boldsymbol{x}|\rho_k)\,d\boldsymbol{x}, \int_{\mathbb{R}^2} V_m(\boldsymbol{x} \mid \rho_m)\,d\boldsymbol{x}\right)\\ &\quad = \rho_k\rho_m \equiv \rho_{km},\ k \neq m.\end{aligned}$$

The value ρ_{km} can be seen as the cross-correlation between the white noise processes V_k and V_m (Ver Hoef and Barry, 1998). Let $f_k,\ k = 1, 2, 3$ be a moving average function defined on $\mathbb{R}^2$, with parameters $\boldsymbol{\theta}_k$. The variable $S_k(\boldsymbol{s}_i)$ is defined by

$$S_k(\boldsymbol{s}_i) = \int_{\mathbb{R}^2} f_k(\boldsymbol{x} - \boldsymbol{s}_i \mid \boldsymbol{\theta}_k)\, V_k(\boldsymbol{x} \mid \rho_k)\,d\boldsymbol{x}.$$

Because the processes V_k are dependent, so are the variables $S_k(\boldsymbol{s}_i), i = 1, \ldots, n$. The conditional distribution of $\boldsymbol{S} = (\boldsymbol{S}_1', \boldsymbol{S}_2', \boldsymbol{S}_3')'$, where $\boldsymbol{S}_j = (S_j(\boldsymbol{s}_1), \ldots, S_j(\boldsymbol{s}_n))'$ is the vector of the latent variable S_j at all sampled locations, is a multivariate Gaussian distribution with mean zero and covariance matrix $\boldsymbol{C}$

$$\boldsymbol{S}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho} \sim \mathcal{N}_{3n}(\boldsymbol{0}, \boldsymbol{C}), \tag{5}$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$. This forms the second level of the hierarchy. One advantage of this construction is that the expression of the covariance matrix $\boldsymbol{C}$ is known

$$\begin{aligned}&\mathrm{Cov}(S_k(\boldsymbol{s}_i), S_m(\boldsymbol{s}_j))\\ &\quad = \rho_{km}\int_{\mathbb{R}^2} f_k(\boldsymbol{x} - \boldsymbol{s}_i \mid \boldsymbol{\theta}_k)\, f_m(\boldsymbol{x} - \boldsymbol{s}_j|\boldsymbol{\theta}_m)\,d\boldsymbol{x},\end{aligned} \tag{6}$$

where $\rho_{kk} \equiv 1$. The value ρ_{km} gives the strength of cross spatial dependence. Depending on the choice of the moving average functions, the calculation of the integral is either explicit or untractable. In the latter case, each element of the matrix can be seen as an autocorrelation in signal theory and can be numerically computed using the Fast Fourier Transform (Ver Hoef, Cressie, and Barry, 2004). Here, the moving average functions were chosen proportional to the Gaussian kernel: $f_k(\boldsymbol{x}|\boldsymbol{\theta}_k) = \sigma_k \exp(-\|\boldsymbol{x}\|^2 / \phi_k)$ and $\boldsymbol{\theta}_k = (\sigma_k, \phi_k)$, which led to an analytical expression for (6)

$$\mathrm{Cov}(S_k(\boldsymbol{s}_i), S_m(\boldsymbol{s}_j)) = \frac{\rho_{km}\,\sigma_k\sigma_m\,\phi_k\phi_m\,\pi}{\phi_k + \phi_m}\exp\left(-\frac{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|^2}{\phi_k + \phi_m}\right).$$

The vector of correlation parameters $\boldsymbol{\rho}$ is not identifiable. For bivariate data sets, only the product $\rho_{12} = \rho_1\rho_2$ can be identified. For K variables ($K > 2$), the K-tuples $(\rho_1, \ldots, \rho_K)$ and $(-\rho_1, \ldots, -\rho_K)$ lead to the same covariance matrix, so the sign of ρ_1 must be fixed to ensure the identifiability of the correlation parameters.

2.3 *Third Level of the Hierarchy*

The third level of the hierarchical model consists in giving the *prior* distributions on the parameters $\mu_k,\ \nu_1, \boldsymbol{\alpha}_3,\ \boldsymbol{\rho}$, and $\boldsymbol{\theta}_k\ (k = 1, 2, 3)$. The *prior* distributions on $\mu_1,\ \mu_2,\ \mu_3$ are uniform distributions. The nugget effect ν_1^2 of the Gaussian variable $Y_1(\boldsymbol{s})$ has an inverse gamma *prior* distribution: $\nu_1^2 \sim \mathrm{IG}(a, b)$, where a and b are sufficiently small to get a non-informative *prior* distribution. An independent uniform *prior* distribution is assigned to each spatial dependence parameter $\boldsymbol{\theta}_k,\ k = 1, 2, 3$ and $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$. The $L-2$ unknown breakpoints $\alpha_{3;\,l},\ l = 2, \ldots, L-1$ related to the ordinal variable are ordered values, so the *prior* distribution of the vector $(\alpha_{3;2}, \ldots, \alpha_{3;L-1})$ is the order distribution of $L-2$ uniform random variables. Web Appendix B specifies these *prior* distributions.

3. Bayesian Implementation

3.1 *Posterior Analysis*

The hierarchical spatial model can be summarized by the *posterior* distribution of the parameters. Using the *prior* distributions, the joint distribution of parameters and latent variables is given by

$$\begin{aligned}&\pi(\mu_1, \mu_2, \mu_3, \boldsymbol{S}, \boldsymbol{Z}_3, \nu_1, \boldsymbol{\alpha}_3, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\rho}|\boldsymbol{Y})\\ &\propto \exp\left\{-\frac{1}{2\nu_1^2}(\boldsymbol{Y}_1 - \mu_1\boldsymbol{1} - \boldsymbol{S}_1)'(\boldsymbol{Y}_1 - \mu_1\boldsymbol{1} - \boldsymbol{S}_1)\right\}\\ &\times \prod_{i=1}^{n}\left[\frac{\{\exp(\mu_2 + S_2(\boldsymbol{s}_i))\}^{Y_2(\boldsymbol{s}_i)}\exp\{-\exp(\mu_2 + S_2(\boldsymbol{s}_i))\}}{Y_2(\boldsymbol{s}_i)!}\right]\\ &\times \prod_{i=1}^{n}\left[\exp\left\{-\frac{1}{2}(Z_3(\boldsymbol{s}_i) - \mu_3 - S_3(\boldsymbol{s}_i))^2\right\}\right.\\ &\left.\times\, \mathbb{1}(Z_3(\boldsymbol{s}_i) \in]\alpha_{3;Y_3(\boldsymbol{s}_i)-1}; \alpha_{3;Y_3(\boldsymbol{s}_i)}])\right]\\ &\times \exp\left\{-\frac{1}{2}\boldsymbol{S}'\boldsymbol{C}^{-1}\boldsymbol{S}\right\}\\ &\times \pi_0(\mu_1)\pi_0(\mu_2)\pi_0(\mu_3)\pi_0(\nu_1^2)\pi_0(\boldsymbol{\alpha}_3)\pi_0(\boldsymbol{\theta}_1)\pi_0(\boldsymbol{\theta}_2)\pi_0(\boldsymbol{\theta}_3)\pi_0(\boldsymbol{\rho})\end{aligned}$$

where $\mathbb{1}$ is the indicator function, $\boldsymbol{1}$ is a vector of length n with all terms equal to 1, $\boldsymbol{Z}_3 = (Z_3(\boldsymbol{s}_1), \ldots, Z_3(\boldsymbol{s}_n))'$ is the vector of the latent variable Z_3 at all sampled locations, and π_0 is the *prior* distribution. A sample from the marginal *posterior* distributions for each of these parameters can be obtained through the implementation of a Markov chain Monte Carlo (MCMC) simulation scheme. The inference procedure is based on a mixture of Gibbs and Metropolis sampling.

Parameters $\mu_1,\ \mu_3,\ \nu_1,\ \boldsymbol{\alpha}_3$, and latent variables $\boldsymbol{S}_1,\ \boldsymbol{S}_3,\ Z_3(\boldsymbol{s}_i)$ for $i = 1, 2, \ldots, n$ are drawn iteratively from their full conditional distribution

$$\mu_1|\ldots \sim \mathcal{N}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_1(\boldsymbol{s}_i)-S_1(\boldsymbol{s}_i)),\frac{\nu_1^2}{n}\right\},$$

$$\mu_3|\ldots \sim \mathcal{N}\left\{\frac{1}{n}\sum_{i=1}^{n}(Z_3(\boldsymbol{s}_i)-S_3(\boldsymbol{s}_i)),\frac{1}{n}\right\},$$

$$\nu_1^2|\ldots \sim \mathrm{IG}\left\{a+\frac{n}{2}, b+\frac{\sum_{i=1}^{n}(Y_1(\boldsymbol{s}_i)-\mu_1-S_1(\boldsymbol{s}_i))^2}{n}\right\},$$

$$Z_3(\boldsymbol{s}_i)|\ldots \sim \mathcal{N}\{\mu_3+S_3(\boldsymbol{s}_i),1\}$$
truncated to $[\alpha_{3;\,Y_3(\boldsymbol{s}_i)-1},\alpha_{3;\,Y_3(\boldsymbol{s}_i)}[$
for $i=1,2,\ldots,n$,

$$\alpha_{3;\,l}|\ldots \sim \mathcal{U}[\max\{\max\{Z_3(\boldsymbol{s}_i):Y_3(\boldsymbol{s}_i)=l\},\alpha_{3;\,l-1}\};$$
$$\min\{\min\{Z_3(\boldsymbol{s}_i):Y_3(\boldsymbol{s}_i)=l+1\},\alpha_{3;\,l+1}\}]$$
for $l=2,\ldots,L-1$,

$$\boldsymbol{S}_1|\ldots \sim \mathcal{N}_n(\boldsymbol{m}_1^*,\boldsymbol{V}_1^*)$$
with $\begin{cases}\boldsymbol{V}_1^*=\left(\boldsymbol{V}_1^{-1}+\frac{1}{\nu_1^2}\boldsymbol{I}_n\right)^{-1}\\ \boldsymbol{m}_1^*=\boldsymbol{V}_1^*\left\{\boldsymbol{V}_1^{-1}\boldsymbol{m}_1+\frac{1}{\nu_1^2}(\boldsymbol{Y}_1-\mu_1\boldsymbol{1})\right\}\end{cases}$

$$\boldsymbol{S}_3|\ldots \sim \mathcal{N}_n(\boldsymbol{m}_3^*,\boldsymbol{V}_3^*)$$
with $\begin{cases}\boldsymbol{V}_3^*=\left(\boldsymbol{V}_3^{-1}+\boldsymbol{I}_n\right)^{-1}\\ \boldsymbol{m}_3^*=\boldsymbol{V}_3^*\left\{\boldsymbol{V}_3^{-1}\boldsymbol{m}_3+(\boldsymbol{Z}_3-\mu_3\boldsymbol{1})\right\},\end{cases}$

where $\mathcal{U}(a,b)$ is the uniform distribution between a and b, $\boldsymbol{I}_n$ is the $n\times n$ identity matrix, and $\boldsymbol{m}_k$ and $\boldsymbol{V}_k$ are the conditional expectation and covariance matrix of $\boldsymbol{S}_k$ given $\boldsymbol{S}_l$, $l\neq k$. These are given by: $\boldsymbol{m}_k=\boldsymbol{\Gamma}_k(\boldsymbol{S}_l',\boldsymbol{S}_m')'$, and $\boldsymbol{V}_k=\boldsymbol{C}_{kk}-\boldsymbol{\Gamma}_k(\boldsymbol{C}_{lk}',\boldsymbol{C}_{m\,k}')'$, where

$$\boldsymbol{\Gamma}_k=\begin{pmatrix}\boldsymbol{C}_{kl} & \boldsymbol{C}_{km}\end{pmatrix}\begin{pmatrix}\boldsymbol{C}_{ll} & \boldsymbol{C}_{lm}\\ \boldsymbol{C}_{ml} & \boldsymbol{C}_{mm}\end{pmatrix}^{-1}$$
$$(l,m=1,2,3;l,m\neq k)$$

and $\boldsymbol{C}_{kl}=\mathrm{Cov}(\boldsymbol{S}_k,\boldsymbol{S}_l)$ is the $n\times n$ covariance matrix that is computed using (6).

The parameter μ_2 and the spatial dependence parameters $\boldsymbol{\theta}_k, k=1,2,3$ and $\boldsymbol{\rho}$ are sampled from a Metropolis step (Hastings, 1970), whereas the latent vector $\boldsymbol{S}_2$ is updated by an adaptive version of the Metropolis Langevin algorithm (Atchade, 2006). Each vector $\boldsymbol{\theta}_k$ and each term of $\boldsymbol{\rho}$ is updated separately. Let $\pi(x)$ be the target distribution of the quantity x. For $\boldsymbol{\theta}_k$ and ρ_k, the target distribution is proportional to $\pi(\boldsymbol{S}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\rho})$ times their *prior* distribution, where $\boldsymbol{S}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\rho}$ is given by (5). The target distribution for $\boldsymbol{S}_2$ is proportional to $\pi(\boldsymbol{Y}_2|\boldsymbol{S}_2,\boldsymbol{\mu}_2)\,\pi(\boldsymbol{S}_2|\boldsymbol{S}_1,\boldsymbol{S}_3,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\rho})$, where $\boldsymbol{Y}_2|\boldsymbol{S}_2,\boldsymbol{\mu}_2$ is given by (2) and $\boldsymbol{S}_2|\boldsymbol{S}_1,\boldsymbol{S}_3,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\rho}\sim\mathcal{N}_n(\boldsymbol{m}_2,\boldsymbol{V}_2)$. A new value x^* is sampled from a proposal distribution $q(.|x)$. The proposal value x^* is accepted with probability $\min\{1,\pi(x^*)q(x|x^*)\,/\,[\pi(x)q(x^*|x)]\}$. The proposal distribution for these parameters is a Gaussian distribution centered on the current value of the parameter, or a truncated Gaussian distribution if there are constraints on the parameter. The proposal distribution for $\boldsymbol{S}_2$ is the n-variate Gaussian distribution with mean $\boldsymbol{S}_2+(\lambda^2/2)\,\mathrm{D}(\boldsymbol{S}_2)$ and covariance matrix $\lambda^2\boldsymbol{I}_n$, where $\lambda>0$ is a variable scale parameter,

$$\mathrm{D}(\boldsymbol{S}_2)=\frac{\delta}{\max(\delta,|\nabla\ln(\pi(\boldsymbol{S}_2))\,|\,)}\nabla\ln(\pi(\boldsymbol{S}_2)),$$

∇ is the gradient operator, and $\delta>0$ is a fixed constant. The scale parameter λ is updated at each iteration of the algorithm in order to obtain a prescribed acceptance rate of 0.574 (Atchade and Rosenthal, 2005), and $\delta=1{,}000$.

3.2 *Prediction*

The goal here is to predict the multivariate random field at n_0 unsampled locations. Let $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{n_0}$ be the n_0 unsampled locations, let $\tilde{\boldsymbol{S}}_j=(S_j(\boldsymbol{u}_1),\ldots,S_j(\boldsymbol{u}_{n_0}))'$ be the n_0-dimensional vector of the latent variable S_j at all unsampled locations, and let $\tilde{\boldsymbol{S}}=(\tilde{\boldsymbol{S}}_1',\tilde{\boldsymbol{S}}_2',\tilde{\boldsymbol{S}}_3')'$. When the updating scheme of the random field parameters has converged, we introduce the following step to generate $\tilde{\boldsymbol{S}}$. At the tth iteration, we draw a $3n_0$-dimensional random sample $\tilde{\boldsymbol{S}}^{(t)}$ from the conditional multivariate Gaussian distribution

$$\tilde{\boldsymbol{S}}^{(t)}|\boldsymbol{S}^{(t)},\boldsymbol{\theta}_1^{(t)},\boldsymbol{\theta}_2^{(t)},\boldsymbol{\theta}_3^{(t)},\boldsymbol{\rho}^{(t)}$$
$$\sim\mathcal{N}_{3n_0}\left(\boldsymbol{C}_{12}'\boldsymbol{C}_{11}^{-1}\boldsymbol{S}^{(t)},\boldsymbol{C}_{22}-\boldsymbol{C}_{12}'\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12}\right),$$

where $\boldsymbol{C}_{11}=\mathrm{Var}(\boldsymbol{S}^{(t)})$, $\boldsymbol{C}_{12}=\mathrm{Cov}(\boldsymbol{S}^{(t)},\tilde{\boldsymbol{S}}^{(t)})$, and $\boldsymbol{C}_{22}=\mathrm{Var}(\tilde{\boldsymbol{S}}^{(t)})$, whose expressions follow from (6). The covariance matrices $\boldsymbol{C}_{11}$, $\boldsymbol{C}_{12}$, and $\boldsymbol{C}_{22}$ are computed using the current values of the parameters $\boldsymbol{\theta}_k^{(t)}, k=1,2,3$ and $\boldsymbol{\rho}^{(t)}$ (Diggle et al., 1998; Kern, 2000; Christensen and Waagepetersen, 2002). Using the current value of $\mu_k^{(t)}, k=1,2,3$, we obtain

- a realization $\tilde{y}_1^{(t)}(\boldsymbol{u}_i)$ of the mean of $Y_1(\boldsymbol{u}_i)$: $\tilde{y}_1^{(t)}(\boldsymbol{u}_i)=\mu_1^{(t)}+\tilde{S}_1^{(t)}(\boldsymbol{u}_i)$. It should be clear that $\tilde{y}_1^{(t)}(\boldsymbol{u}_i)$ is a realization of the smooth process without the sampling error, whereas $\tilde{Y}_1^{(t)}(\boldsymbol{u}_i)\sim\mathcal{N}(\tilde{y}_1^{(t)}(\boldsymbol{u}_i),{\nu_1^{(t)}}^2)$ is a realization of $Y_1(\boldsymbol{u}_i)$ that includes the sampling error;
- a realization $\tilde{y}_2^{(t)}(\boldsymbol{u}_i)$ of the mean of $Y_2(\boldsymbol{u}_i)$: $\tilde{y}_2^{(t)}(\boldsymbol{u}_i)=\exp(\mu_2^{(t)}+\tilde{S}_2^{(t)}(\boldsymbol{u}_i))$. Again it should be clear that $\tilde{y}_2^{(t)}(\boldsymbol{u}_i)$ does not include the sampling error, whereas $\tilde{Y}_2^{(t)}(\boldsymbol{u}_i)\sim\mathcal{P}(\tilde{y}_2^{(t)}(\boldsymbol{u}_i))$ is a realization of $Y_2(\boldsymbol{u}_i)$ that includes the sampling error;
- and a realization $\tilde{z}_3^{(t)}(\boldsymbol{u}_i)$ of the mean of $Z_3(\boldsymbol{u}_i)$: $\tilde{z}_3^{(t)}(\boldsymbol{u}_i)=\mu_3^{(t)}+\tilde{S}_3^{(t)}(\boldsymbol{u}_i)$. A realization $\tilde{y}_3^{(t)}(\boldsymbol{u}_i)$ is then obtained by truncating $\tilde{z}_3^{(t)}(\boldsymbol{u}_i)$ according to the current breakpoints $\boldsymbol{\alpha}_3^{(t)}$.

By letting the MCMC updating scheme run long enough, one can obtain as many realizations $\tilde{y}_k^{(t)}(\boldsymbol{u}_i)$ as desired. For the Gaussian variable, a prediction $\widehat{Y}_1(\boldsymbol{u}_i)$ is the mean of the realizations $\tilde{y}_1^{(t)}(\boldsymbol{u}_i)$. For the count variable, a prediction $\widehat{Y}_2(\boldsymbol{u}_i)$ is the median of the realizations $\tilde{y}_2^{(t)}(\boldsymbol{u}_i)$. We used the median rather than the mean because the mean of $\exp(\mu_2+S_2(\boldsymbol{s}_i))$ may be infinite when a vague prior is used for $\boldsymbol{\theta}_2$ (De Oliveira, Kedem, and Short, 1997). In the same way, a prediction

Table 1

Validation criteria obtained from a simulated data set by univariate, bivariate, and trivariate estimation procedures. For Gaussian and Poisson variables, bias, RMSPE, and RMEV are given. In addition 80 %PI is given for Gaussian variables. The percentage of well-predicted values is given for ordinal variables.

Estimation	Gaussian variable		Poisson variable		Ordinal variable	
Univariate	bias	0.47	bias	−0.34	%CP	76.0
	RMSPE	4.43	RMSPE	2.90		
	RMEV	4.95	RMEV	2.57		
	80%PI	0.81				
Bivariate	bias	0.27	bias	−0.37		
Gaussian–Poisson	RMSPE	4.59	RMSPE	2.75		
	RMEV	4.96	RMEV	2.98		
	80%PI	0.80				
Bivariate	bias	0.38			%CP	79.5
Gaussian-ordinal	RMSPE	4.44				
	RMEV	4.95				
	80%PI	0.83				
Bivariate			bias	−0.34	%CP	80.5
Poisson-ordinal			RMSPE	2.90		
			RMEV	2.58		
Trivariate	bias	0.35	bias	−0.40	%CP	80.0
	RMSPE	4.57	RMSPE	2.78		
	RMEV	5.11	RMEV	2.95		
	80%PI	0.80				

$\widehat{Y}_3(\boldsymbol{u}_i)$ for the ordinal variable is the median of the realizations $\tilde{y}_3^{(t)}(\boldsymbol{u}_i)$.

The accuracy of the predictions is checked by validation. The validation criteria are different according to the type of the variable. Let n_V be the number of sampled locations $\boldsymbol{u}_i$ in the validation data set. Let $\widehat{Y}_1(\boldsymbol{u}_i)$, $1 \leqslant i \leqslant n_V$, be the predicted value of the Gaussian variable at the ith location of the validation data set. Let $\widehat{\mathrm{Var}}(\tilde{Y}_1(\boldsymbol{u}_i))$ be the estimated prediction variance at location $\boldsymbol{u}_i$. The variance $\widehat{\mathrm{Var}}(\tilde{Y}_1(\boldsymbol{u}_i))$ is the sum of the variance of all the realizations $\tilde{y}_1^{(t)}(\boldsymbol{u}_i)$ and of the sampling error approximated by $\widehat{\nu_1}^2$. For the Gaussian variable, we compute the following criteria defined by Ver Hoef et al. (2004)

- $\text{bias} = \frac{1}{n_V}\sum_{i=1}^{n_V}(\widehat{Y}_1(\boldsymbol{u}_i) - Y_1(\boldsymbol{u}_i))$,
- $\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n_V}(\widehat{Y}_1(\boldsymbol{u}_i)-Y_1(\boldsymbol{u}_i))^2}{n_V}}$,
- $\text{RMEV} = \sqrt{\frac{\sum_{i=1}^{n_V}\widehat{\mathrm{Var}}(\tilde{Y}_1(\boldsymbol{u}_i))}{n_V}}$,
- $80\%\text{PI} = \frac{1}{n_V}\sum_{i=1}^{n_V}\mathbb{1}\{\,|\,\widehat{Y}_1(\boldsymbol{u}_i) - Y_1(\boldsymbol{u}_i)| < 1.28\sqrt{\widehat{\mathrm{Var}}(\tilde{Y}_1(\boldsymbol{u}_i))}\}$.

If the estimated prediction variances are correct, then RMEV should be close to RMSPE. The prediction interval coverage 80%PI should be about 80%. For the Poisson variable, bias, RMSPE, and RMEV are computed. For the ordinal variable, the percentage of correctly predicted values (%CP) in the validation data set is used as a validation criterion.

4. Simulations

The performance of the inference algorithm was assessed using simulated data sets. The inference algorithm is time consuming if a large number of variables compose the multivariate random field. Consequently only bivariate or trivariate data sets were simulated. To reduce the burn-in time, the algorithm was first run in the univariate case for each variable, and the estimates thus obtained were used as initial values for the multivariate procedure.

4.1 *Simulated Data Sets*

We here describe the simulation procedure for a trivariate data set consisting of a Gaussian variable Y_1, a Poisson variable Y_2, and an ordinal variable Y_3 with three categories that are identified by the integers 1–3. A simplified similar procedure can produce bivariate data sets. First, 350 locations were randomly chosen in a square $[-10;10]\times[-10;10]$. The vector $\boldsymbol{S}$ of length $3n$ was generated according to a multivariate Gaussian distribution with zero mean and covariance $\boldsymbol{C}$. Then the vector of Gaussian variables $\boldsymbol{Y}_1$ was simulated from a Gaussian distribution $\mathcal{N}_n(\mu_1\mathbf{1} + \boldsymbol{S}_1, \nu_1^2\boldsymbol{I}_n)$. For all $i = 1,\ldots,n$, the variable $Y_2(\boldsymbol{s}_i)$ was obtained by sampling from $\mathcal{P}(\exp(\mu_2 + S_2(\boldsymbol{s}_i)))$. For the ordinal variable, we began simulating the latent vector $\boldsymbol{Z}_3$ from a Gaussian distribution $\mathcal{N}_n(\mu_3\mathbf{1} + \boldsymbol{S}_3, \boldsymbol{I}_n)$. For all $l = 1,\ldots,L$, $Y_3(\boldsymbol{s}_i)$ took the value l if $Z_3(\boldsymbol{s}_i)$ was between the $(l-1)$th and the lth L-quantile of $\boldsymbol{Z}_3$. In our simulations, L was taken equal to three. The data set consisted of 250 locations sampled from the 350 initial ones. The 100 remaining values were used as the validation data set.

4.2 *Simulation Results*

Table 1 compares the validation criteria when using the trivariate, bivariate, or univariate estimation procedures. The bivariate and univariate data sets in this case are the marginal restrictions of the trivariate data set. Additional simulation

results based on bivariate data sets are given in Web Appendix C.

There were no meaningful differences between the validation criteria of the different estimation procedures, whatever the type of variable. The trivariate procedure sometimes did slightly better than the univariate procedure (e.g., bias of the Gaussian variable, percentage of well-predicted values for the ordinal variable), and sometimes did slightly worse (e.g., difference between RMSPE and RMEV for the Gaussian variable, bias of the Poisson variable). Hence, the trivariate estimation procedure brought valid inference, in the sense that it was unbiased for practical purposes, and the prediction intervals were correct.

5. Application: Prediction of Soil Properties

The model was applied on soil data collected in French Guiana to predict soil characteristics.

5.1 *Pedological Data Set*

Data were collected in the Paracou experimental forest in French Guiana (5°15′N, 52°55′W; 0–50 m elevation), 15 km inland from the coast (Gourlet-Fleury, Guehl, and Laroussinie, 2004). The climate is humid tropical with a mean annual rainfall of 2980 mm. The relief consists of a patchwork of hills (100–300 m in diameter and 20–50 m in height) separated by humid valleys. Part of the site is permanently waterlogged.

Soils are mostly acrisoils (FAO-ISRIC-ISSS, 1998) developed over a Precambrian metamorphic formation. The soil is characterized by schists and sandstones and locally crossed by veins of pegmatite, aplite, and quartz. Soil properties were measured in four 250 m × 250 m permanent plots located in the south of the experimental site. The plots were located at some distance from each other and the elevation and slope of these plots were known. Around 70 randomly chosen points were recorded in each plot. A 1.2 m core of soil was extracted at each location for characterization. Soil texture, soil color, and the presence of stones or colored spots were used to classify the soils. Manual perception of clay content and silt dryness was used to distinguish soils exhibiting vertical drainage from soils exhibiting superficial lateral drainage. Six levels of drainage were distinguished to classify varying degrees of hydromorphism. Further details concerning the drainage characteristics can be found in Sabatier et al. (1997).

To apply our model, a trivariate data set was built from the data described above. We focused on slope (Gaussian variable), elevation (ordinal variable), and soil drainage (ordinal variable). The elevation was divided into three classes (<20 m, 20–30 m, >30 m). The soil drainage counted four ordered categories. A total of 327 observations were available and 200 locations were sampled for the estimation. The remaining 127 values were used as the validation data set. Each pair of variables was analyzed separately.

5.2 *Results*

Only results for a Gaussian-ordinal data set (slope and soil drainage) and for an ordinal-ordinal data set (drainage and elevation) are given here. The moving average functions were again Gaussian, but with a different parameterization: ϕ_k and σ_k were replaced by $\phi_k^2/2$ and $\phi_k^{-1}\sqrt{4\sigma_k/\pi}$, respectively. The estimates of parameters are given in Table 2.

Table 2
Estimation of the parameters from bivariate data sets at Paracou. Index 1 *for the parameters refers either to slope (Gaussian variable, with ν_1 as a parameter) or to elevation (ordinal variable, with $\alpha_{1,2}$ as a parameter). Index* 2 *always refers to soil drainage.*

	Slope–drainage		Elevation–drainage	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.
σ_1	17.41	3.38	221.87	59.51
ϕ_1	30.44	4.24	76.04	4.38
ν_1 or $\alpha_{1,2}$	2.59	0.42	19.97	3.58
μ_1	10.58	0.49	10.11	2.85
σ_2	3.08	1.12	4.44	1.89
ϕ_2	76.73	11.73	57.82	9.50
$\alpha_{2,2}$	1.67	0.22	1.87	0.29
$\alpha_{2,3}$	4.39	0.41	5.26	0.71
μ_2	2.88	0.37	3.40	0.53
ρ_{12}	0.16	0.21	−0.80	0.07

As with the simulations (see Web Appendix C), the convergence speed was higher for the parameters related to the Gaussian variable than for those related to the ordinal variables. The estimates obtained for slope (Table 2) were consistent with the range, the sill, and the nugget observed on its empirical variogram. The estimate of the parameter μ_1 for slope was close to the mean slope. Slope and soil drainage were not correlated. The estimates for soil drainage obtained from the slope–drainage data set were consistent with those obtained from the elevation–drainage data set (Table 2). The standard deviation of σ_1 for elevation was high. As expected, a negative correlation between soil drainage and elevation was found: hydromorphic soils coincided with bottomlands. In general, the estimates were not as accurate as in the simulation (compare to Web Table 1). The estimates related to the ordinal variable could be improved by increasing the size of the calibration data set. These results could be explained by the specific spatial pattern of the data. Some additional simulations showed indeed that the accuracy of the estimates decreased when sampled locations were clustered.

The predictions obtained from both data sets are given by Figure 1. Both drainage map and elevation map are consistent with our knowledge of the studied area. Table 3 summarizes the validation criteria. The predictions of slope were slightly biased. RMSPE was close to RMEV, and the prediction interval coverage was perfect, indicating that prediction variance was estimated quite accurately. The percentage of correctly predicted values for drainage using the slope–drainage data set was 68%. Most of the inaccurate predictions of the ordinal variable (drainage) were located near the boundaries of plots or far from neighbors, which coincided with locations where the prediction variance for the Gaussian variable (slope) was high. Moreover, the number of observations per category for the ordinal variable were unbalanced. The lack of information about some categories may explain some mistakes in the predictions. The same phenomena were observed for the predictions of drainage using the elevation–drainage data set, where only 67% of values were correctly predicted. Elevation did better with 79% of correctly predicted values.

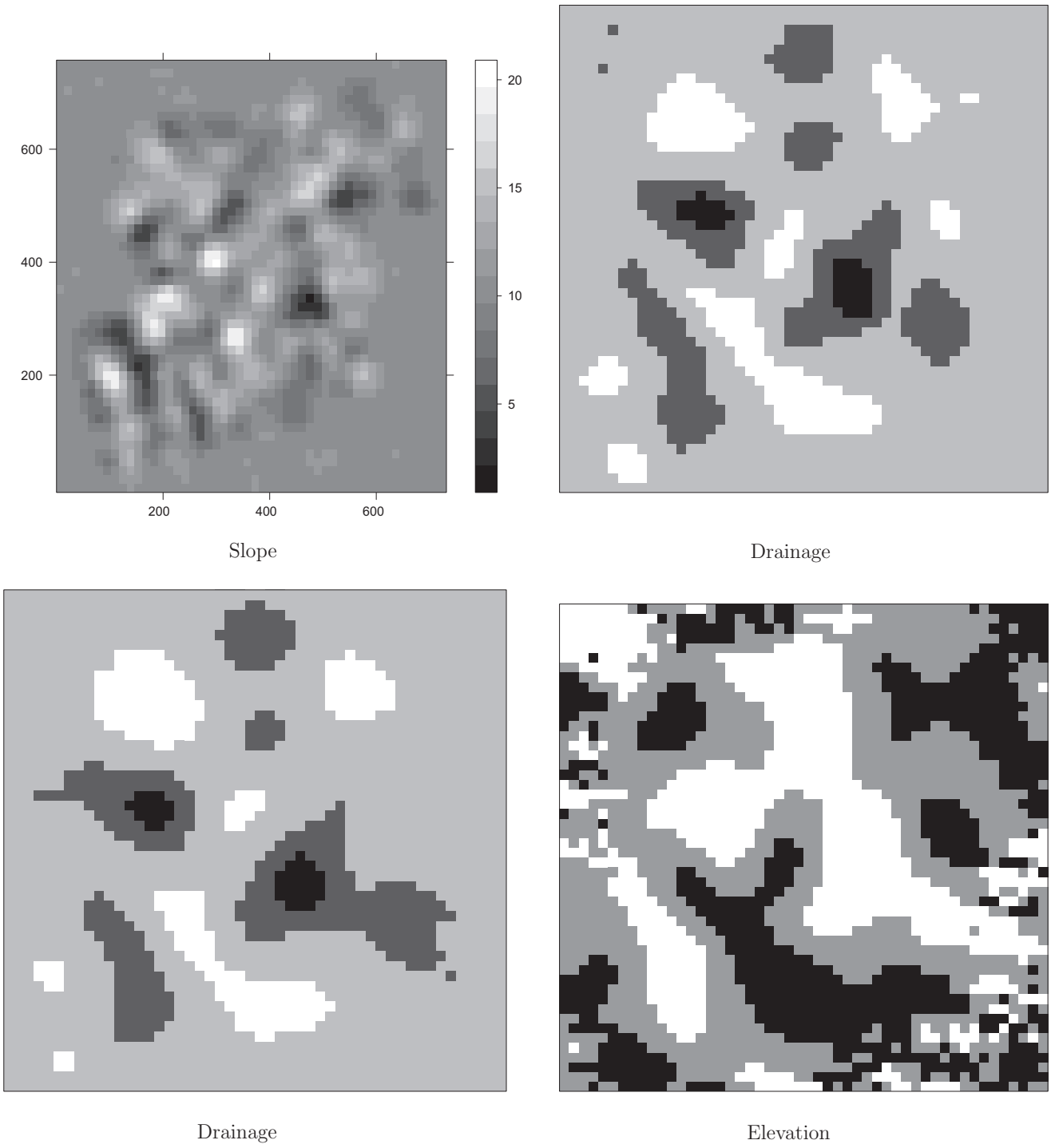


Figure 1. Prediction maps obtained from slope–drainage data set (top) and from drainage–elevation data set (bottom). Slope is given in degrees. Levels of drainage are coded on a grayscale from black (well-drained soils) to white (hydromorphic soils). Levels of elevation are coded on a grayscale from black (low elevation) to white (high elevation). Each rectangle corresponds to an area 747 m $\times$ 765 m.

6. Discussion and Conclusion

A multivariate spatial model to simultaneously predict different types of variables was defined. This model is hierarchical and can deal with Gaussian, Poisson, and ordinal variables using a unified approach based on general linear mixed models. Ordinal variables were addressed using the distribution that follows from the multivariate ordinal probit model. This approach had been widely used as a generalization of

Table 3
Validation criteria for the prediction of slope, elevation, and soil drainage at Paracou

Data set	Slope or elevation		Drainage	
Gaussian-ordinal	bias	−0.68	%CP	67.7
	RMSPE	4.91		
	RMEV	4.27		
	80%PI	0.80		
Ordinal-ordinal	%CP	78.7	%CP	66.9

Euclidian distance for mixed continuous and discrete data (Bedrick, Lapidus, and Powell, 2000; Mortier et al., 2006; Chaubert et al., 2008). Recently, in a spatial context, Augustin et al. (2007) used this model with nonlinear effects of covariates and spatial random effects to predict ordinal variables. However, Augustin et al. (2007) is a regression-based approach that requires knowing the values of covariates at unsampled locations to predict the ordinal variable at these locations. On the contrary, our approach predicts the multivariate random field without requiring the values of covariates to be known.

6.1 *Limits of the Model*

Although the model was able to predict different types of variables while taking account of their correlations, some problems also came into view. First, the inference method seems to be sensitive to the spatial pattern of sampled locations and to edge effects. Spatial clustering of sampled locations would lead to a poorer representation of the spatial structure, and then to a poorer quality of predictions than when sampled locations are located randomly or regularly. This may explain why the quality of predictions was poorer with pedological data than with simulated data. We may also suspect from the prediction of elevation and drainage at Paracou that the accuracy of predictions for ordinal data decreases with the number of categories in the ordinal variable.

A particular moving average function had to be chosen for inference on the real data set. This choice is crucial because the form and the flexibility of the variogram that translates the dependence between variables depend on it. The Gaussian kernel that we chose is attractive because of its limited number of parameters and the simple evaluation of the integrals in equation (6). The moving average functions must be square integrable. It may be possible to use a piecewise constant function as recommended by Ver Hoef et al. (2004). This kind of function leads to a very flexible covariance, but many parameters have to be estimated unless they are constrained by a functional relationship. An alternative would consist of taking some functions that are less flexible but more parsimonious, like disk kernels constructed by stacking cylinders with smaller and smaller radii (Kern, 2000). How to choose and validate a particular function f_k remains an open problem. As a first step, it would be interesting to test the robustness of the model according to the moving average function chosen.

6.2 *Computations*

The inference procedure can become computationally intensive when the number of variables or the number of observations increases because of the size of the covariance matrix in this case. Some alternative methods based on composite marginal likelihoods (Varin, 2008) could be considered. The inference procedure could be simplified following Joe's (1997) approach. The inference method would consist in running as many univariate procedures as the number of variables to estimate the parameters related to each variable, followed by multivariate inference to estimate the vector of correlations $\boldsymbol{\rho}$ and make the predictions. Another solution would be to use approximate Bayesian inference (Eidsvik, Martino, and Rue, 2009; Rue, Martino, and Chopin, 2009) instead of MCMC simulations.

6.3 *Extensions*

While constructing the model, we made several simplifying choices that can be relaxed to extend the model. First, the mixture of white noise processes that we chose led to a symmetric covariance matrix $\boldsymbol{C}$, in the sense that $\mathrm{Cov}(S_k(-\boldsymbol{s}_i), S_m(-\boldsymbol{s}_j)) = \mathrm{Cov}(S_m(\boldsymbol{s}_i), S_k(\boldsymbol{s}_j))$. It would be possible to add a spatial shift in the mixture of white noise processes V_k to introduce a shift-asymmetry of cross spatial dependence (Ver Hoef and Barry, 1998). Another extension would consist of replacing the white noise processes by other spatial processes such as Lévy processes.

One can also relax the simplifying assumption that consists of taking the correlation matrix $\boldsymbol{R}$ equal to the identity matrix when the model comprises several ordinal variables. The model can be extended if the latent Gaussian variables $Z_k(\boldsymbol{s})$ related to the ordinal variables are not independent given $S_k(\boldsymbol{s})$, as in the definition given in Web Appendix A. Finally, an extension of the model can be considered for nominal variables. In the same way as we have generalized the ordinal probit model to deal with ordinal variables, we can generalize the multinomial probit model to take nominal variables into account.

7. Supplementary Materials

Web Appendices A, B, and C and Web Table 1 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors thank the associate editor and two anonymous referees for very useful comments that improved the presentation of the article.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88,** 669–679.

Atchade, Y. F. (2006). An adaptive version for the Metropolis Adjusted Langevin Algorithm with a truncated drift. *Methodology and Computing in Applied Probability* **8,** 235–254.

Atchade, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11,** 815–828.

Augustin, N. H., Lang, S., Musion, M., and von Wilpert, K. (2007). A spatial model for the needle losses of pine-trees in the forest of Baden-Württemberg: An application of Bayesian structured additive regression. *Applied Statistics* **56,** 29–50.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, Florida: Chapman and Hall/CRC.

Barry, R. P. and Ver Hoef, J. M. (1996). Blackbox kriging: Spatial prediction without specifying variogram models. *Journal of Agricultural Biological and Environmental Statistics* **1,** 297–322.

Bedrick, E. J., Lapidus, J., and Powell, J. F. (2000). Estimating the Mahalanobis distance from mixed continuous and discrete data. *Biometrics* **56,** 394–401.

Calder, C. A. and Cressie, N. (2007). Some topics in convolution-based spatial modeling. In *Proceedings of the 56th Session of the ISI, August 22-29, 2007*, Voorburg, The Netherlands. International Statistical Institute.

Chaubert, F., Mortier, F., and Saint-André, L. (2008). Multivariate dynamic model for ordinal outcomes. *Journal of Multivariate Analysis* **99,** 1717–1732.

Chen, M.-H. and Shao, Q.-M. (1999). Properties of prior and posterior distributions for multivariate categorical response data models. *Journal of Multivariate Analysis* **71,** 277–296.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85,** 347–361.

Chica-Olmo, J. (2007). Prediction of housing location price by a multivariate spatial method: Cokriging. *Journal of Real Estate Research* **29,** 92–114.

Christensen, O. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using Generalized Linear Mixed Models. *Biometrics* **58,** 280–286.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing* **6,** 101–111.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: John Wiley & Sons.

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics, and Data Analysis* **34,** 299–314.

De Oliveira, V., Kedem, B., and Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association* **92,** 1422–1433.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47,** 299–350.

Eidsvik, J., Martino, S., and Rue, H. (2009). Approximate Bayesian inference in spatial generalized linear mixed models. *Scandinavian Journal of Statistics* **36,** 1–22.

FAO-ISRIC-ISSS. (1998). *World Reference Base for Soil Resources*. Rome: Food and Agricultural Organization of the United Nations.

Gelfand, A. E., Banerjee, S., Sirmans, C. F., Tu, Y., and Ong, S. E. (2007). Multilevel modeling using spatial processes: Application to the Singapore housing market. *Computational Statistics, and Data Analysis* **51,** 3567–3579.

Golam Kibria, B. M., Sun, L., Zidek, J. V., and Le, N. D. (2002). Bayesian spatial prediction of random space-time fields with application to mapping $PM_{2.5}$ exposure. *Journal of the American Statistical Association* **97,** 112–124.

Gourlet-Fleury, S., Guehl, J. M., and Laroussinie, O. (2004). *Ecology and Management of Neotropical Rainforest: Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*. Paris: Elsevier.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109.

Higdon, D. (2001). Space and space-time modeling using process convolutions. Duke Statistics Discussion Papers 2001-03, Duke University, Durham, North Carolina.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall/CRC.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Applied Statistics* **52,** 1–18.

Kern, J. C. (2000). Bayesian process-convolution approaches to specifying spatial dependence structure. Ph.D. Thesis, Duke University, Durham, North Carolina.

Matheron, G. (1965). *Les variables régionalisées et leur estimation*. Paris, France: Masson.

Matheron, G. (1976). A simple substitute for conditional expectation: The disjunctive kriging. In *Advanced Geostatistics in the Mining Industry*, M. Guarascio, M. David, and C. Huijbregts (eds), 221–236. Reidel, The Netherlands: Dordrecht.

McBratney, A. B., Odeh, I. O. A., Bishop, T. F. A., Dunbar, M. S., and Shatar, T. M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma* **97,** 293–327.

Molchanov, I. (1997). *Statistics of the Boolean Model for Practitioners and Mathematicians*. Chichester, U.K.: John Wiley & Sons.

Mortier, F., Robin, S., Lassalvy, S., Baril, C., and Bar-Hen, A. (2006). Prediction of Euclidian distances with discrete and continuous outcomes. *Journal of Multivariate Analysis* **97,** 1799–1814.

Rivoirard, J. (1991). *Introduction au krigeage disjonctif et à la géostatistique non linéaire*, 2nd edition. Fontainebleau: École Nationale Supérieure des Mines de Paris.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* **71,** 319–392.

Sabatier, D., Grimaldi, M., Prevost, M. F., Guillaume, J., Godron, M., Dosso, M., and Curmi, P. (1997). The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. *Plant Ecology* **131,** 81–108.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis* **92,** 1–28.

Ver Hoef, J. M. and Barry, R. P. (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference* **69,** 275–294.

Ver Hoef, J. M., Cressie, N., and Barry, R. P. (2004). Flexible spatial models for kriging and cokriging using moving averages and the Fast Fourier Transform (FFT). *Journal of Computational and Graphical Statistics* **13,** 265–282.

Wackernagel, H. (2003). *Multivariate Geostatistics. An Introduction with Applications*, 3rd edition. Berlin: Springer Verlag.

Webster, R. and Oliver, M. A. (1990). *Statistical Methods in Soil and Land Resource Survey*. Oxford: Oxford University Press.

Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecology processes. *Ecology* **84,** 1382–1394.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85,** 251–267.

Received February 2009. *Revised January* 2010.
Accepted February 2010.

Population dynamics of species-rich ecosystems : the mixture of matrix population models approach

Methods in Ecology and Evolution 2013

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 316–326 doi: 10.1111/2041-210x.12019

Population dynamics of species-rich ecosystems: the mixture of matrix population models approach

Frédéric Mortier[1]*, Vivien Rossi[2], Gilles Guillot[3], Sylvie Gourlet-Fleury[1] and Nicolas Picard[1]

[1]*CIRAD,UPR Bsef, Montpellier, 34398, France;* [2]*CIRAD, UMR Ecofog, Kourou, 97387, Guyane, France; and* [3]*Statistics Section IMM, Technical University of Denmark, Copenhagen, Denmark*

Summary

1. Matrix population models are widely used to predict population dynamics, but when applied to species-rich ecosystems with many rare species, the small population sample sizes hinder a good fit of species-specific models. This issue can be overcome by assigning species to groups to increase the size of the calibration data sets. However, the species classification is often disconnected from the matrix modelling and from the estimation of matrix parameters, thus bringing species groups that may not be optimal with respect to the predicted community dynamics.

2. We proposed here a method that jointly classified species into groups and fit the matrix models in an integrated way. The model was a special case of mixture with unknown number of components and was cast in a Bayesian framework. An MCMC algorithm was developed to infer the unknown parameters: the number of groups, the group of each species and the dynamics parameters.

3. We applied the method to simulated data and showed that the algorithm efficiently recovered the model parameters.

4. We applied the method to a data set from a tropical rain forest in French Guiana. The mixture matrix model classified tree species into well-differentiated groups with clear ecological interpretations. It also accurately predicted the forest dynamics over the 16-year observation period.

5. Our model and algorithm can straightforwardly be adapted to any type of matrix model, using the life cycle diagram. It can be used as an unsupervised classification technique to group species with similar population dynamics.

Key-words: Bayesian, clustering, mixture models, reversible jump Markov chain Monte Carlo, tropical rain forests, species-rich ecosystems, population dynamics

Introduction

The conservation of animal and plant species and their biological control require models to understand and predict population dynamics (Fieberg & Ellner 2001; Buongiorno & Gilless 2003; Demyanov, Wood & Kedwards 2006). Among population dynamics models, projection matrix models have been widely used to investigate the dynamics of age-, stage- or size-structured populations (Caswell 2001; Stott *et al.* 2010). They provide a simple way of integrating vital rate information such as recruitment, birth, growth or ageing, and mortality (Crone *et al.* 2011). Matrix models have been used to model population demography in the context of species invasion (Hooten *et al.* 2007; Sebert-Cuvillier *et al.* 2007), species extinction or conservation of endangered species (Cropper & Loudermilk 2006), and the sustainable management of exploited species (Hauser, Cooch & Lebreton 2006). Recent improvements in matrix models targeted the estimation of demographic parameters, in particular for animal populations using capture–recapture methods (Besbeas *et al.* 2002).

In species-rich ecosystems like tropical rain forests, tropical marine fish or coral reefs, high diversity implies that the sample size for most species is limited. The small sample size hinders a good fit of species-specific dynamics models, including matrix population models. To address this problem, modellers usually cluster species into groups. A variety of methods has been used to group species, favouring either ecological interpretation or the accuracy of predictions. Groups of species can be derived from functional characteristics (Steneck & Dethier 1994), ecomorphology (Bellwood & Wainwright 2001) or ecological subjective strategy (Swaine & Whitmore 1988; Favrichon 1994; Gitay & Noble 1997). These methods do not rely on a strong statistical methodology, thus they do not ensure that the within-group similarity is maximum, or that the number of groups is optimal. Gourlet-Fleury *et al.* (2005) described two other strategies applied in tropical rain forests: the ecological data-driven strategy (Phillips *et al.* 2002) and the dynamic process strategy, in which 'process' refers to the components of forest dynamics (recruitment, growth or mortality)

*Correspondence author. E-mail: frederic.mortier@cirad.fr

(Gourlet-Fleury & Houllier 2000; Picard *et al.* 2010). These strategies rely on statistical unsupervised classification methods, such as hierarchical cluster analysis, to group species with similar traits. Moreover, species classification is most often disconnected from the matrix modelling and from the estimation of the matrix parameters, thus bringing species groups that may not be optimal with respect to the predicted community dynamics. To cluster the species while ensuring optimality for predicting community dynamics, we need to rely on the mixture model framework.

Mixture models are based on the assumption that observation data arise from several unobserved groups (McLachlan & Peel 2000). A model is associated to each group. Each observation contributes to the fitting of the model for a given group with a weight that represents its probability to belong to this group. These weights can eventually be used to classify observations among groups. Thus, mixture modelling simultaneously fits models and classifies observations, and the clustering step is closely linked to the calibration step. This favours the similarity of species response within groups rather than the similarity of species traits (Dunstan, Foster & Darnell 2011). Mixture modelling has mainly been developed for observations with a normal distribution (e.g. mixture regressions). The use of mixture models has recently been proposed to model the presence/absence of species (Dunstan, Foster & Darnell 2011), the species richness in a species assemblage (Mao, Colwell & Chang 2005) or the heterogeneity of capture and survival probabilities in free-ranging populations (Pledger, Pollock & Norris 2010).

This study aims at extending mixture modelling to matrix population models. The mixture of matrix population models will simultaneously solve two issues: fit matrix models for species-rich ecosystem with many rare species, and classify species into groups. As proposed in population genetics (Pritchard, Stephens & Donnelly 2000; Corander, Waldmann & Sillanpaa 2003; Guillot *et al.* 2005), the strategy consists in a probabilistic model-based clustering method expressed in terms of matrix population mixture models with an unknown number of components (Richardson & Green 1997; Dunson 2000; Marin, Mengersen & Robert 2005). The number of groups and the parameters of the matrix population models associated with each group are the unknown quantities. We propose to use a Bayesian framework to infer these unknown quantities. The Bayesian framework has several advantages over frequentist methods. First, it enables us to obtain the credibility interval for finite population sizes, whereas frequentist methods provide asymptotic confident intervals. Secondly, with the use of prior distributions, strong biological or ecological knowledge can be integrated in the model.

The mixture of matrix models is defined in the next section. An inference method is then outlined, and tested using simulated data. The mixture matrix model is finally applied to a data set from the Paracou tropical rain forest in French Guiana. The tree species groups obtained had consistent ecological behaviours with contrasted functional traits, and compared favourably to other groups obtained by a standard classification technique.

Materials and methods

MIXTURE OF MATRIX POPULATION MODELS

When fitting a base model to some observations, it is assumed that the set of observations is homogeneous, in the sense that all observations share a common distribution (e.g. the normal distribution for the residuals of a linear model). When dealing with an heterogeneous set of observations composed of K assumedly homogeneous subsets, mixture modelling is a relevant framework to extend this base model (McLachlan & Peel 2000). Mixture model assumes that the distribution of observations is a mixture of K base distributions, with mixing weights that represent the probability for an observation to belong to each of the homogeneous subsets. Conditionally on an observation belonging to a subset, the model identifies with the base model, while the distribution of the mixture includes the uncertainty on which subset an observation belongs to.

Mixture of matrix population models results from the application of the mixture framework to matrix population models. In matrix population models, individuals are classified into stage, size or age classes, and the population dynamics is described by transition rates among classes (Caswell 2001). At the individual level, these transitions can be interpreted as the transitions of a Markov chain, which defines some distribution of the population-level numbers of individuals that switched between two classes. Assuming that individuals have any of K, such dynamics distribution defines a mixture of K matrix population models. A specificity of the mixture of matrix models is that one observation corresponds to one population (more specifically, it is the vector of all numbers of individual transitions between classes), and the set of observations is the community-level set of populations. Hence, mixtures of matrix models are relevant to model the dynamics of a community when assuming that its constituent species can be assigned to K homogeneous groups of species.

Hereafter, we detail the mathematical expression of the mixture of matrix models for a specific type of matrix population models, namely the Usher model. This framework readily extends to any type of matrix models on the basis of individual transitions among classes.

MIXTURE OF USHER MATRIX MODELS

The Usher matrix model applies to size-structured populations (Usher 1966, 1969). It is based on the description of the change of the population by a vector, $\vec{N}_t$ containing the numbers $N_{l,t}$ of individuals in L ordered size classes ($l = 1,\ldots, L$) at discrete time t. Let $N_t = \sum_{l=1}^{L} N_{l,t}$ be the total number of individuals at time t. Like any other matrix population model, the Usher model can be interpreted as the expectation of N_t independent Markov chains (Fig. 1). The relationship between $\vec{N}_t$ and $\vec{N}_{t+1}$ is described by a $L \times L$ transition matrix U, called the Usher matrix:

$$\mathrm{E}[\vec{N}_{t+1}|\vec{N}_t] = U\mathrm{E}[\vec{N}_t] \qquad \text{eqn 1}$$

where:

$$U = \begin{pmatrix} p_1 + f & f & \cdots & f \\ q_1 & p_2 & & 0 \\ & \ddots & \ddots & \\ 0 & & q_{L-1} & p_L \end{pmatrix} \qquad \text{eqn 2}$$

and p_l is the probability for an individual to stay in class l, q_l the probability to move up from class l to $l + 1$ and f the average fecundity. q_l and p_l take values in [0, 1], whereas f takes positive real values. The probability to die for an individual in class l is given by

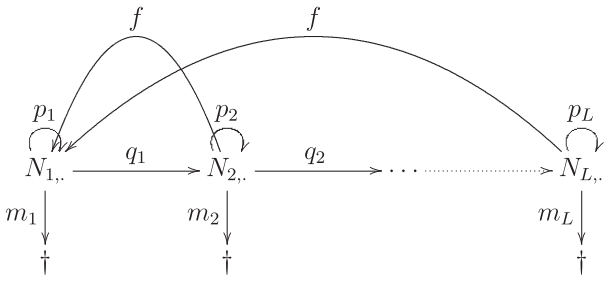


Fig. 1. Life cycle representation of the Usher projection matrix model, where *pl* is the probability for an individual to stay in class *l*, q_l is the probability to move up from class *l* to *l* + 1, m_l is the probability of dying and *f* is the average fecundity.

$m_l = 1 - p_l - q_l$. Let $\vec{d} = (d_1, \ldots, d_L)$ be the class distribution of the population, such that d_l denotes the probability for a randomly chosen individual to belong to class *l* ($\sum_{l=1}^{L} d_l = 1$). Let $N_{l,l,t}$ denote the number of individuals staying in class *l* between $t - 1$ and *t*, $N_{l,l+1,t}$ the number of individuals moving up from class *l* to $l + 1$ between $t - 1$ and *t*, and $N_{l,\dagger,t}$ the number of individuals dying in class *l* between $t - 1$ and *t*. Let R_t be the number of recruits between $t - 1$ and *t*, assumed to be a Poisson random variable with parameter $f\,N_{t-1}$. The vector of observations for the population is $\vec{N} = (N_{1,l,t}, \ldots, N_{L,\dagger,t}, \vec{N}_{t-1}, R_t)$. The likelihood of the joined individual Markov transitions, and thus of the Usher matrix model, is:

$$\begin{aligned}\mathcal{L}(\vec{N}|\theta) = \prod_{l=1}^{L-1} & \mathcal{M}(N_{l,l,t}, N_{l,l+1,t}, N_{l,\dagger,t}|p_l, q_l, m_l, N_{l,t-1}) \\ & \times \mathcal{M}(N_{L,L,t}, N_{L,\dagger,t}|p_L, m_L, N_{L,t-1}) \\ & \times \mathcal{M}(N_{1,t-1}, \ldots, N_{L,t-1}|d_1, \ldots, d_L, N_{t-1}) \\ & \times \mathcal{P}(R_t|fN_{t-1})\end{aligned} \qquad \text{eqn 3}$$

where $\mathcal{M}$ denotes the multinomial distribution, $\mathcal{P}$ the Poisson distribution and $\theta = (\vec{p}, \vec{q}, \vec{m}, f, \vec{d})$ is the vector of parameters with $\vec{p} = (p_1, \ldots, p_L)$, $\vec{q} = (q_1, \ldots, q_{L\text{-}1})$ and $\vec{m} = (m_1, \ldots, m_L)$. Equation 1 is the deterministic version of the Usher projection model while eqn 3 accounts for the demographic stochasticity and is useful when the population size gets small Caswell 2001.

Suppose now that the modelled population arises from *K* unobserved groups of species such that each group is modelled by a Usher matrix model. Thus, there are *K* Usher matrices $U_1, \ldots, U_K$. Because the group the population belongs to is not known a priori, one can define a random latent variable *C* that identifies the group of the species. For example, if the species belongs to the third group: conditionally on $C = 3$, the prediction of the dynamics is given by eqn 1, with *U* being replaced by U_3. Accounting for the uncertainty on *C* brings:

$$\mathrm{E}[\vec{N}_{t+1}|\vec{N}_t] = \sum_{k=1}^{K} \pi_k U_k \mathrm{E}[\vec{N}_t] \qquad \text{eqn 4}$$

where π_k is the posterior probability that *C* equals *k*. Equation 4 defines the mixture of Usher matrix models, whose likelihood is:

$$\mathcal{L}(\vec{N}|\vec{\theta}, \vec{\pi}) = \sum_{k=1}^{K} \pi_k \mathcal{L}(\vec{N}|\theta_k) \qquad \text{eqn 5}$$

where $\vec{\theta} = (\theta_1, \ldots, \theta_K)$ is the vector of all parameters associated with the *K* matrix models, $\vec{\pi} = (\pi_1, \ldots, \pi_K)$ is the vector of all posterior probabilities, and $\mathcal{L}(\vec{N}|\theta_k)$ is given by eqn 3. The species can be a posteriori classified by assigning it to the group *g* with the maximum posterior probability: $\pi_g = \max_k\{\pi_k\}$. Hence, the mixture of matrix models jointly defines *K* matrix models (and implicitly provides us with a way to estimate $\vec{\theta}$) and classifies the species into *K* groups (i.e. provides an estimate of $\vec{\pi}$).

MIXTURE MODEL INFERENCE

The parameters $\vec{\theta}$ and $\vec{\pi}$ of the mixture matrix model can be estimated in a frequentist context by maximizing the likelihood (5) of the mixture model. Inference can be achieved using an EM algorithm (McLachlan & Krishnan 2008). However, we used the Bayesian inference framework to have the opportunity to integrate biological knowledge into the model through the prior distribution of the parameters. Based on the direct acyclic graph of the mixture matrix model (Fig. 2), a Markov chain Monte Carlo (MCMC) inference algorithm was implemented: a long sequence of parameter values was randomly drawn from the posterior distribution, and the parameter estimates were extracted from this sample by computing its mode or its means (Gilks, Richardson & Spiegelhalter 1996). Details on the Bayesian inference, including the choice of the priors, are given in Appendix A. Annotated R codes (R Core Team 2012) for the algorithm and a first tentative version of MPMM package are available in the Supporting Information.

Fitting a mixture model also requires estimating the number *K* of groups. Classically, different mixture models with different numbers of groups are independently fitted, and an information criterion is finally used in the end to perform selection among these competing models (Biernacki, Celeux & Govaert 2000 ; see also Cubaynes et al. 2012 in a capture-recapture context). A MCMC algorithm for a fixed *K* was developed with this aim in view. Alternatively, we also developed an inference algorithm that considered *K* as unknown and jointly estimated it with the other parameters. This involved using a reversible jump MCMC approach when the number of groups changed (Richardson & Green 1997). With this latter approach, posterior probabilities for each value of *K* were obtained, thus enabling one to choose the most likely *K* while assessing the reliability of this choice.

Because the posterior distribution for the number *K* of groups may be sensitive to changes in the prior distribution for of the parameters when using a reversible jump MCMC algorithm (Richardson & Green

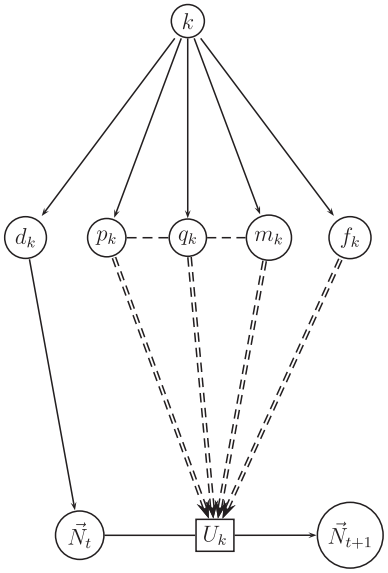


Fig. 2. Direct acyclic graph of the mixture of Usher projection matrix model. Double dot arrows indicate deterministic links, dot lines indicate direct links, circles indicate random nodes and frames indicate deterministic nodes.

1997), a sensitivity analysis to the priors was performed. Details on the different priors that were tested are given in Appendix A.

SIMULATIONS

Data were simulated to assess the efficiency of the algorithm to correctly classify species into groups, according to different levels of differentiation between groups and different numbers of groups. Simulated data were composed of 100 species distributed across eight diameter classes. Numerical experiments tested the combinations of three factors: (i) the number of groups, that was equal to 1, 5 or 10 (three modalities), and will be referred to as the true number of groups; (ii) the number of individuals per species, that was equal to 100 or 1000 (two modalities); and (iii) hyper-priors for parameters ($\vec{d}$, $\vec{p}$, $\vec{q}$, $\vec{m}$, f), that took the values given in Table 1 (five modalities).

The five different hyper-priors for the parameters corresponded to five levels of differentiation between groups. Indeed, the expectation of the diameter class or transition parameters was constant ($E(d_l) = 1/8$ and $E(p_l) = E(q_l) = E(m_l) = 1/3$ for all the hyper-priors in Table 1), but their variances decreased from 0·012 to 0·0015 for d_l and from 0·055 to 0·0079 for the transition parameters. As this variance corresponded to the between-group variance, the lower it was, the more similar the groups were. Let us note Ldiff_1,…, Ldiff_5, the five decreasing differentiation levels of the hyper-parameters. When the number of groups was one, only the level Ldiff_1 was used for hyper-priors. In total, there were thus: $2 \times 1 + 2 \times 2 \times 5 = 22$ combinations of factors in the numerical experiments. For each combination, 50 replications were simulated. For each replication, the 100 species were randomly assigned to groups. This simulated classification was the reference to compare with the estimated classification and was referred as the 'true classification'. Then, for each group, the diameter class parameters, the transition parameters and the fecundity parameter were randomly drawn according to their hyper-prior distributions (Table 1). Finally, for each species, the prescribed number of individuals was drawn according to the law defined by eqn 3 using the parameters of the group to which the species belonged.

To assess the performance of the method, we compared the estimated number $\hat{K}$ of groups with the true number K used to simulate data sets, and we compared the estimated classification with the true classification using two set matching indices I_1 and I_2. These indices are based on the $K \times \hat{K}$ contingency table $T = (T_{ij})$ with $i = 1, \ldots, K$ and $j = 1, \hat{K}$ that cross-*tabulates* the species according to the true and the estimated classifications:

$$I_1 = \frac{1}{S}\sum_{i=1}^{K} \max\{T_{i1}, \ldots, T_{i\hat{K}}\} \quad \text{and} \quad I_2 = \frac{1}{S}\sum_{j=1}^{\hat{K}} \max\{T_{1j}, \ldots, T_{Kj}\}$$

These indices vary between $1/S$ and 1, and the higher they are, the better is the adequacy between the two classifications (Meilă 2007). They jointly reflect how groups collapsed and merged: $I_1 = 1$ and $I_2 = 1$ means that both classifications were identical; $I_1 = 1$ and $I_2 < 1$ means that the number of groups was underestimated and one or more groups were merged; $I_1 < 1$ and $I_2 = 1$ means that the number of groups was overestimated and one or more groups were split; $I_1 < 1$ and $I_2 < 1$ means that several set operations are needed to move from one classification to the other.

TROPICAL FOREST DATA

Data on the tropical rain forest were collected at the Paracou experimental site (5°18′N, 52°53′W), French Guiana. The site is located in a undisturbed *terra firme* forest under equatorial climate. Three 250 × 250 m permanent sample plots (18·75 ha in total) have been established in 1984 and left as control of the undisturbed forest dynamics. All trees greater than 10 cm d.b.h have been identified and georeferenced. Girth at breast height, standing deaths, treefalls and newly recruited trees greater than 10 cm d.b.h have been monitored either annually or every 2 years since 1984 (Gourlet-Fleury, Guehl & Laroussinie 2004). Because the Paracou forest is a mature undisturbed forest, the diameter distribution in those control plots could be considered at quasi-equilibrium. Two data sets were extracted from the Paracou data base: one training data set to infer the mixture of Usher models, and one validation data set. A data set gave the species, the diameter class at year t and the diameter class at year $t + 2$ for n trees. Trees that died between years t and $t + 2$, and trees whose diameter overcame the inventory threshold of 10 cm between years t and $t + 2$ (recruited individuals) were included in the data set.

The training data set consisted of the data collected in 1993 and 1995 on the three control plots. One hundred and eighty-one species were identified in these three control plots (Fig. 3), illustrating both the high species richness, and the relative scarcity of most species of the Guianan forest. The mean number of individuals per species was 64·54 (total on the three control plots of the training data set), with a minimum of 1 and a maximum of 980. The median number of individuals per species was 22, with a first quartile of 8 and a third quartile of 61·25. Although it could be possible to include species with few individuals into the analysis, we decided to leave out species with less than 20 individuals in the control plots in 1993. A preliminary analysis (data not shown) evidenced that there was little difference between the classification based on all species and the classification restricted to species having at least 20 individuals: the algorithm took longer to converge in the former case, rare species were not well classified, and actually behaved like noise with respect to the estimation of groups. Moreover, from an ecological point of view, it does not make sense to assign species to groups when they are represented by few individuals. It is ecologically much more meaningful to a posteriori assign rare species to existing groups, using expert's knowledge on the species autecology. Hence, we reckon that rare species should rather be a posteriori assigned to existing

Table 1. Hyper-prior distributions of the parameters used for simulations. $\mathcal{D}$ is the Dirichlet distribution, $\mathcal{G}$ is the gamma distribution. 'Var' is the variance of d_i, of p_l, q_l, m_l, and of f respectively

	Diameter $\vec{d}$		Transition (p_l, q_l, m_l)		Fecundity f	
Differentiation level	Distribution	Var	Distribution	Var	Distribution	Var
Ldiff_1	$\mathcal{D}$(1,1,1,1,1,1,1,1)	0·0121	$\mathcal{D}$ (1,1,1)	0·055	$\mathcal{G}$ (10, 1000)	10^{-5}
Ldiff_2	$\mathcal{D}$ (3,3,3,3,3,3,3,3)	0·0044	$\mathcal{D}$ (3,3,3)	0·022	$\mathcal{G}$ (10, 2000)	$2{\cdot}5 \times 10^{-6}$
Ldiff_3	$\mathcal{D}$ (5,5,5,5,5,5,5,5)	0·0027	$\mathcal{D}$ (5,5,5)	0·014	$\mathcal{G}$(10, 3000)	$1{\cdot}1 \times 10^{-6}$
Ldiff_4	$\mathcal{D}$ (7,7,7,7,7,7,7,7)	0·0019	$\mathcal{D}$ (7,7,7)	0·010	$\mathcal{G}$ (10, 4000)	$6{\cdot}25 \times 10^{-7}$
Ldiff_5	$\mathcal{D}$ (9,9,9,9,9,9,9,9)	0·0015	$\mathcal{D}$ (9,9,9)	0·008	$\mathcal{G}$ (10, 5000)	4×10^{-7}

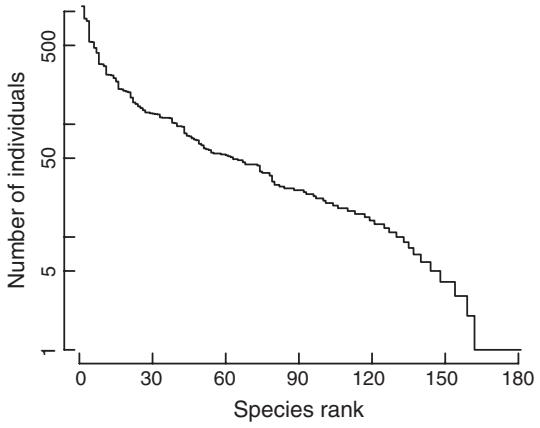


Fig. 3. Rank-abundance diagram in the control plots at Paracou in 1993.

groups. We were left with 93 species that included at least 20 trees monitored in the three control plots. This training data set contained 10 756 trees. The validation data set consisted of the data collected in 2009 on the same three control plots.

A classification of tree species into five groups was defined at Paracou by Favrichon (1994) using multivariate analysis and *k*-means clustering of species attributes (including size summary statistics, growth and recruitment). On the basis of these groups, Favrichon (1998) then fitted a Usher matrix model to predict forest dynamics. Hence, Favrichon's approach is illustrative of a two-step approach with a species classification that is disconnected from the matrix population model. We compared Favrichon's species classification with the one obtained by the mixture matrix model using the likelihood (5) of the training data set. Because there were missing observations between 1995 and 2009, the same computation was intractable for the validation data set. Nevertheless, considering that the undisturbed forest was close to equilibrium, we also compared the likelihoods of the validation data set given the asymptotic diameter distributions according to the two classifications. For a given population with Usher transition matrix U (eqn 2), the asymptotic diameter distribution is the normalized eigenvector of U associated to its dominant eigenvalue (Caswell 2001).

Results

RECOVERY OF SIMULATED CLASSIFICATIONS

Simulation results were similar whether we used a uniform or a truncated Poisson distribution as a prior for K. Hence, only the results with the later prior (that was the default one) are reported here. For 1000 individuals per species, the estimated classification perfectly matched with the true simulated classification for all differentiation levels: I_1 and I_2 were always equal to one.

For 100 individuals per species, the results depended on the differentiation levels and on the number of groups (Table 2). When the true number of groups was one, the algorithm always found one group. For five groups, we correctly estimated the number of groups in 100%, 100%, 96%, 76% and 52% of the cases for the five decreasing levels of differentiation respectively. When the number of groups was wrongly estimated, it was systematically underestimated: I_1 was very close to 1 and I_2 always remained lower than I_1. The classification method tended to merge different species groups into one group, and to dispatch very few species of a given group into another group. The same results were found with stronger evidence in the case of 10 groups. At the fourth level of differentiation, the number of group was correctly estimated in about 80% of the cases, and more than 95% of the species were classified into the correct groups.

Table 2. Comparison between simulated and estimated classifications: mean of (I_1, I_2) on the 50 simulations for 100 individuals per species, depending of the differentiation levels for the hyper-priors. Definition of Ldiff_i is given in Table 1

Differentiation level	One group	Five groups	Ten groups
Ldiff_1	(1, 1)	(1, 1)	(1, 1)
Ldiff_2	n.d.	(0·996, 0·996)	(0·998, 0·988)
Ldiff_3	n.d.	(0·996, 0·989)	(0·978, 0·889)
Ldiff_4	n.d.	(0·983, 0·933)	(0·929, 0·686)
Ldiff_5	n.d.	(0·964, 0·865)	(0·899, 0·574)

n.d., not defined.

TROPICAL RAIN FOREST TREE SPECIES CLASSIFICATION

The 93 tree species at Paracou were classified using the mixture of matrix models, based on eight diameter classes (≤ 15 cm, 15–20, 20–25, 25–30, 30–40, 40–50, 50–60, ≥ 60 cm). Based on 50 different chains, and 20 000 iterations after a burn-in of 10 000 iterations, five groups were obtained 48 times and six groups twice. Groups remained globally the same for all chains. We kept the chain with the highest log-likelihood. For this chain, the posterior probabilities for $K = 5$, 6, 7 or 35 groups were equal to 0·99, $5{\cdot}3 \times 10^{-3}$, $9{\cdot}3 \times 10^{-4}$ and $6{\cdot}7 \times 10^{-5}$ respectively.

The sensitivity analysis to the prior distributions showed that the estimate of K was fairly insensitive to the specification of the prior distributions for the parameters. For all priors except one, the algorithm found again five groups of species. The exception corresponded to $\alpha = \beta = 10$ for the priors of the transition and diameter class parameters, to be compared to $\alpha = \beta = 1$ for the default prior (Appendix A). In that case, K was estimated to three groups (with former groups 2 and 3 merged into a single one, and former groups 4 and 5 merged into a single one). Because α and β can be interpreted as pseudo-counts of individuals in diameter classes, large values of α and β tend to decrease the impact of observations on the classification, in particular for the largest diameter class that have few observations. Hence, the sensitivity of K to α and β expresses the sensitivity of the species classification to differences between species in the largest diameter classes.

To help interpreting the five species groups, five demographic and biological attributes were computed for each group: growth rate, mortality rate, fecundity rate, upper bound for diameter and turnover. Direct estimates of these attributes were computed from the training data set, and compared to the indirect estimates obtained from the estimated transition

and diameter class parameters of the mixture matrix model (see the Supporting Information for the estimates of all mixture matrix model parameters). The direct estimate of growth was the mean diameter increment between 1993 and 1995 of all trees that belonged to the group, while its indirect estimate was $\sum_{i=1}^{L-1} p_i d_i \delta_i$, where δ_i is the width of the ith diameter class. The direct estimate of the mortality was the ratio of the number of dead trees in the group between 1993 and 1995 over the number of trees in the group in 1993, while its indirect estimate was $\sum_{i=1}^{L} m_i d_i$. The direct estimate of the fecundity was the ratio of the number of recruited trees in the group between 1993 and 1995 over the number of trees in the group in 1993, while its indirect estimate was f. The direct estimate of the upper bound for diameter was the 95% quantile of diameters in 1995, while its indirect estimate was interpolated from $\vec{d}$ assuming that the diameter distribution was uniform within each class. Finally, the turnover was computed as half the sum of the mortality rate and of the fecundity rate. The direct and indirect estimates of these attributes were not expected to be strictly equal since they did not derive from the same estimators; yet, their values were quite close and evidenced the same differences between groups (Table 3).

Groups were labelled by decreasing order of growth (Table 3). The gradients of maximum size and turnover perfectly paralleled this gradient of growth, with the fastest growing group 1 having the greatest maximum size and the lowest turnover rate. Group 1 was composed of emergent mid-tolerant species, i.e. species that need to settle in the upper strata and sometimes above the forest canopy to complete their whole life cycle. Group 2 was composed of a mix of shade-tolerant (mostly) and light-demanding (to a lesser extent) canopy species. Group 3 was composed of shade-tolerant species, with a mix of canopy (mostly) and understorey (to a lesser extent) species. As a consequence, its growth rate and maximum size were lower than for group 2, but higher than for group 4. The two small-sized groups 4 and 5 were composed of understorey shade-tolerant species, although group 4 also included a few pioneer species. As a consequence, the growth rate of group 4 was higher than that of group 5.

Because mixture of matrix models jointly classifies species and fits matrix models, we also compared the predicted and the observed number of individuals in each diameter class and each group in 2009, to check the validity of the matrix model. The mixture matrix population model correctly predicted both the number of trees 16 years later and their size distribution (Fig. 4).

The log-likelihood of the training data set was −2722·7 for the Bayesian classification and −3351·7 for Favrichon's classification. The log-likelihood of the validation data set given the asymptotic diameter distribution was −2007·7 for the Bayesian classification and −2874·3 for Favrichon's classification. Hence, both criteria largely favoured the Bayesian classification to the detriment of Favrichon's classification.

Discussion

Mixture modelling can deal with matrix population models, and can jointly classify species and fitting matrix models. Mixture of matrix population models can be addressed in the frequentist or in the Bayesian context. The algorithm that we developed in the Bayesian context performed well on simulated data with known groups, even when the differentiation between groups was low. Classification was correctly predicted when between-group variances were higher than 0·0019 for diameter parameters ($\vec{d}_k$) and 0·010 for transition parameters ($\vec{p}_k, \vec{q}_k, \vec{m}_k$ and f_k), corresponding to the fourth level of differentiation (see Table 1). A specificity of the Bayesian method presented here is that it estimated the number K of groups together with the other parameters. This is original as mixture modelling generally operates conditionally on K, and then uses an information criterion to select K (Biernacki, Celeux & Govaert 2000). Moreover, the Bayesian approach allowed us to construct prior distributions taking into account ecological expert knowledge. For example, we assumed that the prior diameter distribution was a Dirichlet distribution where all parameters were equal to one meaning that the diameter distribution was uniform across diameter classes. Nevertheless, using the Bayesian paradigm, it is straightforward to change the prior distribution to model expert knowledge, assuming for example that the diameter distribution is decreasing from the first to the last diameter class.

The method that we developed for the mixture of Usher matrix models could straightforwardly be adapted to other types of matrix projection models, such as Leslie or Lefkovitch matrix models for age- and stage-structured populations respectively. Starting from the life cycle representation of the matrix model (Fig. 1), one simply has to translate the probabil-

Table 3. Observed vital rates of groups (Obs.) and average vital rates computed from the estimated transition rates (Est.): 2-year d.b.h increment (ΔDBH), 2-year mortality rate, 2-year fecundity rate, upper bound of diameters (DBH95) and 2-year turnover of the five groups obtained using matrix population mixture model classification. The observed ΔDBH for group i was $\frac{1}{k_i}\sum_{j=1}^{k_i}(Y_j^{1995} - Y_j^{1993})$, where Y_j^t was the d.b.h of individual j at year t, and k_i the number of individuals in group i

	ΔDBH (cm)		Mortality (%)		Fecundity (%)		DBH95 (cm)		Turnover (%)	
Group	Obs.	Est.	Obs.	Est.	Obs.	Est.	Obs.	Est.	Obs.	Est.
1	0·38	0·42	0·91	1·31	1·25	1·25	65·3	68·1	1·08	1·28
2	0·27	0·25	1·33	1·58	1·04	1·05	44·2	45·6	1·19	1·32
3	0·24	0·24	2·34	2·70	1·02	1·09	37·4	37·8	1·68	1·90
4	0·13	0·10	2·21	2·38	1·54	1·47	24·2	24·7	1·87	1·93
5	0·08	0·05	2·18	2·74	1·86	2·03	16·4	17·9	2·02	2·39

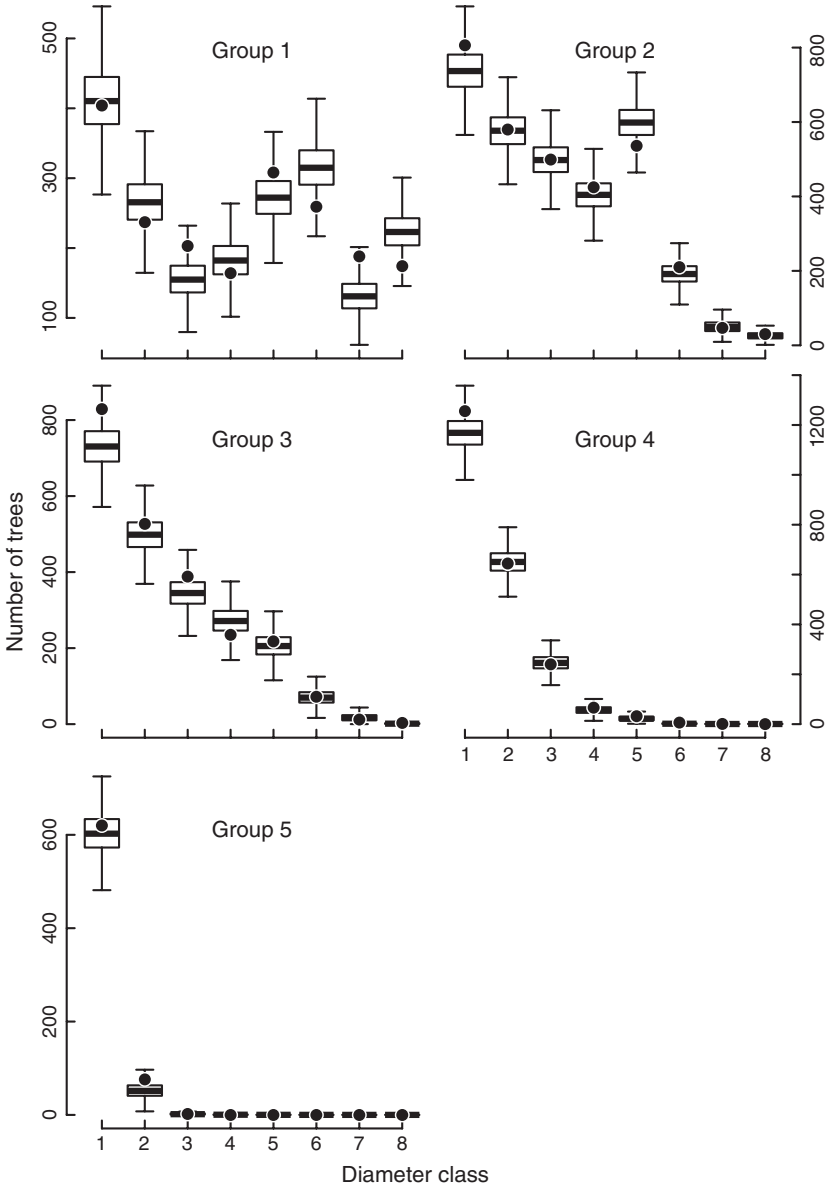


Fig. 4. Predicted (boxplot) and observed (black dot) number of individuals in each diameter class and each species group in the control plots at Paracou in 2009.

ities associated to each transition into a distribution law for an observation (eqn 3).

When applied to a tropical rain forest at Paracou, the mixture of Usher matrix models was able to jointly classify species and make reliable predictions. Predictions were better with the mixture model than with Favrichon's two-step approach, thus exemplifying that a classification disconnected from the matrix model may not be optimal to predict the community dynamics. The characteristics of the tree species groups formed at Paracou were consistent with known ecological behaviour (Lieberman *et al.* 1985; Nascimento *et al.* 2005; Delcamp *et al.* 2008; Poorter *et al.* 2008): small-sized species (with the exception of pioneers) tend to grow slowly, to have high recruitment and mortality rates (i.e. high turnover rates), whereas large sized species that reach the forest canopy tend to grow rapidly and have low turnover rates. The mixture of Usher matrix models classified species according to both their growth rate and their maximum size (Picard *et al.* 2012). When plotting species along these two axes, species groups were clearly separated (Fig. 5). Because these two axes can be used to order species along a continuum of ecological strategies (Turner 2001; Alder *et al.* 2002), this means that the mixture of Usher matrix models was also able to classify species in a way that is consistent with their autecology.

The heterogeneity, in terms of light-requirement, found in groups 2 and 4 can be easily understood given the environmental conditions prevailing in the control plots. These plots are largely undisturbed, with only small gaps occurring at a rate of more or less 3 per year (Gourlet-Fleury, Guehl & Laroussinie 2004). Such conditions do not favour the growth of light-demanding species, nor the growth and survival of pioneer species. Because these species do not express their

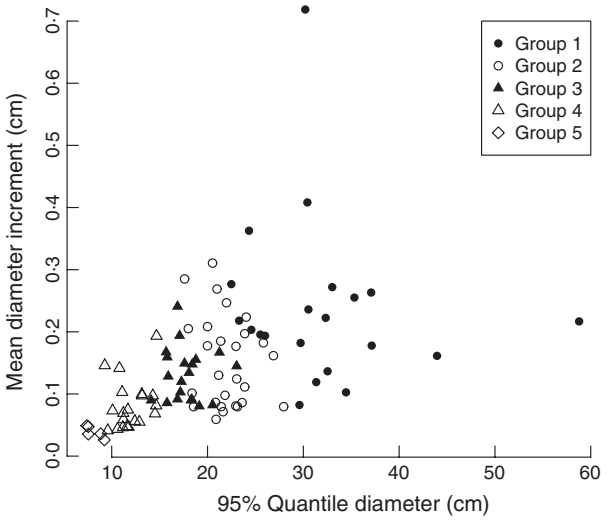


Fig. 5. Upper bound of diameters (95% quantile of d.b.h. in 1995, in cm) vs. mean diameter increment between 1993 and 1995 (cm) for 93 species at Paracou, French Guiana. The five different symbols correspond to the five groups defined by the mixture matrix model.

growth potential, they tended to be gathered with slower growing species in groups 2 and 4. This, in addition to the fact that few pioneer species can survive in these plots, explains why no pioneer group was identified by our procedure while such a group usually is the first one to be isolated in a classification, due to its particular behaviour (Swaine & Whitmore 1988). Applying the mixture of matrix models to disturbed plots would have raised a different classification better accounting for the variety of potential specific behaviours.

In the Paracou example, the distribution of individuals across diameter classes in 1993 was taken into account in the mixture of matrix models: the likelihood (eqn 3) depended on the vector of parameters $\vec{d}$. This means that the shape of the initial diameter distribution influenced the outcome of the species classification. This made sense for the Paracou control plots because these plots were settled in undisturbed forest, whose state in 1993 could be considered as close to equilibrium. The vector $\vec{d}$ was thus representative of the equilibrium state of the forest. We checked indeed (results not shown here) that the asymptotic growth rate of the matrix models was close to one, and the associated eigenvectors close to $\vec{d}$. In other situations where the forest is far from equilibrium, it might not be advisable to account for the initial diameter distribution $\vec{d}$ in the species classification. Computing the conditional likelihood knowing $\vec{N}_t$ would enable to drop $\vec{d}$ from the expression of the likelihood (eqn 3). Apart from this, the mixture of matrix models would be unchanged.

Acknowledgements

This study is part of the GUYASIM project (31032, operational program FEDER 2007–2013), with financial support from European structural funds. This work also has benefited from an 'Investissement d'Avenir' grant managed by the Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alder, D., Oavika, F., Sanchez, M., Silva, J.N.M., Van der Hout, P. & Wright, H.L. (2002) A comparison of species growth rates from four moist tropical forest regions using increment-size ordination. *International Forestry Review*, **4**, 196–205.

Atwood, C.L. (1996) Constrained noninformative priors in risk assessment. *Reliability Engineering and System Safety*, **53**, 37–46.

Bellwood, D. & Wainwright, P. (2001) Locomotion in labrid fishes: implications for habitat use and cross-shelf biogeography on the great barrier reef. *Coral Reefs*, **20**, 139–150.

Besbeas, P., Freeman, S.N., Morgan, B.J.T. & Catchpole, E.A. (2002) Integrating mark–recapture–recovery and census data to estimate animal abundance and demographic parameters. *Biometrics*, **58**, 540–547.

Biernacki, C., Celeux, G. & Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

Buongiorno, J. & Gilless, J.K. (2003) *Decision Methods for Forest Resource Management*. Academic Press, Amsterdam, The Netherlands.

Caswell, H. (2001) *Matrix Population Models: Construction, Analysis and Interpretation*, 2nd edn. Sinauer, Sunderland, Massachusetts, USA.

Corander, J., Waldmann, P. & Sillanpaa, M.J. (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Crone, E.E., Menges, E.S., Ellis, M.M., Bell, T., Bierzychudek, P., Ehrlén, J., Kaye, T.N., Knight, T.M., Lesica, P., Morris, W.F., Oostermeijer, G., Quintana-Ascencio, P.F., Stanley, A., Ticktin, T., Valverde, T. & Williams, J. (2011) How do plant ecologists use matrix population models? *Ecology Letters*, **14**, 1–8.

Cropper, W.P.J. & Loudermilk, E.L. (2006) The interaction of seedling density dependence and fire in a matrix population model of longleaf pine (*Pinus palustris*). *Ecological Modelling*, **198**, 487–494.

Cubaynes, S., Lavergne, C., Marboutin, E. & Gimenez, O. (2012) Assessing individual heterogeneity using model selection criteria: How many mixture components in capture-recapture models? *Methods in Ecology and Evolution*, **3**, 564–573.

Delcamp, M., Gourlet-Fleury, S., Flores, O. & Gamier, E. (2008) Can functional classification of tropical trees predict population dynamics after disturbance? *Journal of Vegetation Science*, **19**, 209–220.

Demyanov, V., Wood, S.N. & Kedwards, T.J. (2006) Improving ecological impact assessment by statistical data synthesis using process-based models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **55**, 41–62.

Dunson, D.B. (2000) Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society*, **62**, 335–336.

Dunstan, P.K., Foster, S.D. & Darnell, R. (2011) Model based grouping of species across environmental gradient. *Ecological Modelling*, **222**, 955–963.

Favrichon, V. (1994) Classification des espèces arborées en groupes fonctionnels en vue de la réalisation d'un modèle de dynamique de peuplement en forêt guyanaise. *Revue d'Écologie (Terre et Vie)*, **49**, 379–403.

Favrichon, V. (1998) Modeling the dynamics and species composition of tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes. *Forest Science*, **44**, 113–124.

Fieberg, J. & Ellner, S.P. (2001) Stochastic matrix models for conservation and management: a comparative review of methods. *Ecology Letters*, **4**, 244–266.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds. (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Gitay, H. & Noble, I.R. (1997) What are functional types and how should we seek them? *Plant Functional Types: Their Relevance to Ecosystem Properties and Global Change* (eds. T.M. Smith, H.H. Shugart & F.I. Woodward), pp. 3–19. International Geosphere-Biosphere Programme. Cambridge University Press, Cambridge, UK.

Gourlet-Fleury, S., Guehl, J.M. & Laroussinie, O., eds. (2004) *Ecology and Management of a Neotropical Rainforest. Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*. Elsevier, Paris, France.

Gourlet-Fleury, S. & Houllier, F. (2000) Modelling diameter increment in a lowland evergreen rain forest in French Guiana. *Forest Ecology and Management*, **131**, 269–289.

Gourlet-Fleury, S., Cornu, G., Jésel, S., Dessard, H., Jourget, J.G., Blanc, L. & Picard, N. (2005) Using models for predicting recovery and assessing tree species vulnerability in logged tropical forests: a case study from French Guiana. *Forest Ecology and Management*, **209**, 69–85.

Guillot, G., Estoup, A., Mortier, F. & Cosson, J.F. (2005) A spatial statistical model for landscape genetics. *Genetics*, **170**, 1261–1280.

Hauser, C.E., Cooch, E.G. & Lebreton, J.D. (2006) Control of structured populations by harvest. *Ecological Modelling*, **196**, 462–470.

Hooten, M.B., Wikle, C.K., Dorazio, R.M. & Royle, J.A. (2007) Hierarchical spatiotemporal matrix models for characterizing invasions. *Biometrics*, **63**, 558–567.
Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A*, **186**, 453–461.
Lieberman, D., Lieberman, M., Hartshorn, G.S. & Peralta, R. (1985) Growth rates and age-size relationships of tropical wet forest trees in Costa Rica. *Journal of Tropical Ecology*, **1**, 97–109.
Mao, C.X., Colwell, R.K. & Chang, J. (2005) Estimating the species accumulation curve using mixtures. *Biometrics*, **61**, 433–441.
Marin, J-M., Mengersen, K. & Robert, C.P. (2005) Bayesian modelling and inference on mixtures of distributions. *Bayesian Thinking, Modeling and Computation* (eds. D. Dey & C.R. Rao), pp. 459–507. Handbook of Statistics. Elsevier, Amsterdam, The Netherlands.
McLachlan, G.J. & Krishnan, T. (2008) *The EM Algorithm and Extensions*, 2nd edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, USA.
McLachlan, G. & Peel, D. (2000) *Finite Mixture Models. Wiley Series in Probability and Statistics*. Wiley-Interscience, New York City, New York, USA.
Meilă, M. (2007) Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, **98**, 873–895.
Nascimento, H.E.M., Laurance, W.F., Condit, R., Laurance, S.G., D'Angelo, S. & Andrade, A.C. (2005) Demographic and life-history correlates for amazonian trees. *Journal of Vegetation Science*, **16**, 625–634.
Nobile, A. (2005) *Bayesian finite mixtures: a note on prior specification and posterior computation*. Technical Report 05-3, University of Glasgow.
Phillips, P.D., Yasman, I., Brash, T.E. & van Gardingen, P.R. (2002) Grouping tree species for analysis of forest data in Kalimantan (Indonesian Borneo). *Forest Ecology and Management*, **157**, 205–216.
Picard, N., Mortier, F., Rossi, V. & Gourlet-Fleury, S. (2010) Clustering species using a model of population dynamics and aggregation theory. *Ecological Modelling*, **221**, 152–160.
Picard, N., Köhler, P., Mortier, F. & Gourlet-Fleury, S. (2012) A comparison of five classifications of species into functional groups in tropical forests of French Guiana. *Ecological Complexity*, **11**, 75–83.
Pledger, S., Pollock, K.H. & Norris, J.L. (2010) Open capture-recapture models with heterogeneity. *II: Jolly-Seber model. Biometrics*, **66**, 883–890.
Poorter, L., Wright, S.J., Paz, H., Ackerly, D.D., Condit, R., Ibarra-Manrquez, G., Harms, K.E., Licona, J.C., Martnez-Ramos, M., Mazer, S.J., Muller-Landau, H.C., Peña-Claros, M., Webb, C.O. & Wright, I.J. (2008) Are functional traits good predictors of demographic rates? Evidence from five neotropical forests. *Ecology*, **89**, 1908–1920.
Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Richardson, S. & Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.
Robert, C.P. & Casella, G. (2005) *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York City, New York, USA.
Sebert-Cuvillier, E., Paccaut, F., Chabrerie, O., Endels, P., Goubet, O. & Decocq, G. (2007) Local population dynamics of an invasive tree species with a complex life-history cycle: a stochastic matrix model. *Ecological Modelling*, **201**, 127–143.
Steneck, R. & Dethier, M. (1994) A functional-group approach to the structure of algal dominated communities. *Oikos*, **69**, 476–498.
Stott, I., Townley, S., Carslake, D. & Hodgson, D.J. (2010) On reducibility and ergodicity of population projection matrix models. *Methods in Ecology and Evolution*, **1**, 242–252.
Swaine, M.D. & Whitmore, T.C. (1988) On the definition of ecological species groups in tropical rain forests. *Vegetatio*, **75**, 81–86.
Turner, I.M. (2001) *The Ecology of Trees in the Tropical Rain Forest*. Cambridge Tropical Biology Series, Cambridge University Press, Cambridge, UK.
Usher, M.B. (1966) A matrix approach to the management of renewable resources, with special reference to the selection forests. *Journal of Applied Ecology*, **3**, 355–367.
Usher, M.B. (1969) A matrix model for forest management. *Biometrics*, **25**, 309–315.

Received 8 August 2012; accepted 8 November 2012
Handling Editor: Dr. Olivier Gimene Z

Appendix A

Bayesian inference

Let S be the number of species in the calibration data set. Using the same notation as in the section 'Mixture of Usher matrix models' with the additional superscript s, let $\vec{N}^s = (N^s_{1,l,t},\ldots, N^s_{L,\dagger,t}, \vec{N}^s_{t-1}, R^s_t)$ be the vector of observations for species $s = 1, \ldots, S$ and let $\underline{N} = (\vec{N}^1, \ldots, \vec{N}^S)$ be the vector of observations for all species. Let $\vec{C} = (C_1, \ldots, C_S)$ be the latent vector that gives the group of each species. Considering K as unknown, the posterior probability π_k follows from the posterior density distribution of the mixture model:

$$\pi^N_{\vec{C},\vec{\theta},K}(\vec{C},\vec{\theta},K|\underline{N}) \propto \prod_{s=1}^{S} \mathcal{L}(\vec{N}^s|\theta_{C_s})\ \pi^0_{\vec{C}|\vec{\theta},K}(\vec{C}|\vec{\theta},K)\ \pi^0_{\vec{\theta}|K}(\vec{\theta}|K)\ \pi^0_K(K) \tag{6}$$

where $\mathcal{L}(\vec{N}^s|\theta_{C_s})$ is given by eqn 3, and $\pi^0_{\vec{C}|\vec{\theta},K}$, $\pi^0_{\vec{\theta}|K}$ and π^0_K are the prior densities associated with the class latent random variables, the parameters of each matrix model and the number of groups respectively. For full Bayesian inference of the model, we set the followings priors on the unknown quantities $\vec{C}$, $\vec{\theta}$ and K.

We assumed that the prior distribution for the number K was a Poisson distribution with mean one, truncated to strictly positive values: $\pi^0_K(K) \equiv \mathcal{P}(1)\backslash\{0\}$. This prior distribution was suggested by Nobile (2005) to be more parsimonious than under uniform distribution. For the sensitivity analysis, a uniform distribution between one and S was also used as a prior for K.

The parameters associated with the matrix population model for group k are $(\vec{p}_k, \vec{q}_k, \vec{m}_k), f_k$ and $\vec{d}_k$. The prior for the parameters $\vec{\theta}$ of the K matrix population models assumed that the parameters of the different classes and groups were independent:

$$\pi^0_{\vec{\theta}|K}(\vec{\theta}|K) = \prod_{k=1}^{K}\{\prod_{l=1}^{L-1} \pi^0_{p,q,m|l,k}(p_{lk}, q_{lk}, m_{lk})\}\ \pi^0_{p,m|k}(p_{Lk}, m_{Lk})\ \pi^0_{\vec{d}|k}(\vec{d}_k)\ \pi^0_{f|k}(f_k)$$

Because the Dirichlet distribution (denoted $\mathcal{D}$) is the conjugate prior of the multinomial distribution, we used the Dirichlet distribution as a prior for all transition parameters and all diameter class parameters: $\pi^0_{\vec{d}|k} \equiv \mathcal{D}(\alpha,\ldots,\alpha)$, $\pi^0_{p,q,m|l,k} \equiv \mathcal{D}(\beta,\beta,\beta)$ and $\pi^0_{p,m|k} \equiv \mathcal{D}(\beta,\beta)$, where α and β are hyper-parameters that can be interpreted as pseudo-counts of individuals. The default priors used $\alpha = \beta = 1$. For the sensibility analysis, we also tested $\alpha = \beta = 0{\cdot}5$ that corresponds to the non-informative Jeffreys prior (Jeffreys 1946; Atwood 1996), and $\alpha = \beta = 10$. Because the gamma distribution (denoted $\mathcal{G}$) is the conjugate prior of the Poisson distribution, we used the gamma distribution as a prior for the fecundity parameter: $\pi^0_{f|k} \equiv \mathcal{G}(\gamma,\delta)$, where δ and γ are hyper-parameters. The default prior used $\gamma = 0{\cdot}01$ and $\delta = 1$, which expresses the expert's

knowledge that the recruitment rate in undisturbed natural rain forest is around 1%. For the sensitivity analysis, we also tested $\gamma = 0{\cdot}5$ and $\delta = 1$, 10^{-1} or 10^{-10} (but the Jeffreys prior that corresponds to $\gamma = 0{\cdot}5$ and $\delta = 0$ could not be used because it is improper).

The prior for the class vector $\vec{C}$ assumed that, given the number of groups, each species could equally and independently of the other species be in any group: $\pi^0_{\vec{C}|\vec{\theta},K}(\vec{C}|\vec{\theta},K) = \prod_{s=1}^{S} \pi^0_{\vec{C}|K}(C_s|K)$ where $\pi^0_{\vec{C}|K}(C_s|K)$ is a uniform distribution on the number of groups: $\mathcal{U}(1,\ldots,K)$.

The inference of parameters was made through the investigation of the posterior distribution $\pi^{\underline{N}}_{\vec{C},\vec{\theta},K}(\vec{C},\vec{\theta},K|\underline{N})$ defined by eqn 6. As the number of groups was unknown, the posterior distribution was not available in an analytic form. Hence, a specific Metropolis within Gibbs MCMC algorithm was developed. The algorithm consisted of three moves: increasing the number of groups (birth case); decreasing the number of groups (death case); keeping the same number of groups, but potentially changing one species assignment (no jump case). In the first two cases, the number of parameters was not constant, so a reversible jump MCMC approach was used (Richardson & Green 1997), whereas in the third case, a Gibbs step could be used. All moves were equally distributed with probability 1/3.

In the following, we detail the proposal step for the three moves and the selection step for the birth and death cases.

1. Proposal step. Let $|k|$ denote the number of species in group k, for $k = 1, \ldots, K$. Let K^* denote the number of groups of the proposal and $\vec{C}^\star$ denote the latent class vector of the proposal.

• No jump case: $K^* = K$. The proposal $\vec{C}^\star = (C_1^\star, \ldots, C_S^\star)$ for the latent class vector is drawn in two steps:

(a) randomly choose one species s among the groups that include two or more species;

(b) new assignment $C_s^\star$ for species s is sampled from a multinomial distribution $\mathcal{M}(1; w_1, , w_K)$, whereas $C_t^\star = C_t$ for $t \neq s$. The coefficients w_k are equal to $w_k = \frac{\mathcal{L}(\vec{N}^s|\theta_k)}{\sum_{j=1}^{K}\mathcal{L}(\vec{N}^s|\theta_j)}$ where $\mathcal{L}$ is given by (3).

• Birth case: $K^* = K + 1$. The proposal for the latent class vector is obtained by splitting one group into two subgroups:

(a) randomly choose one group k among the groups that include two or more species; this group will form two subgroups labelled k_1 and k_2;

(b) choose the number $|k_1|$ of species that will compose group k_1 following a uniform distribution: $|k_1| \sim \mathcal{U}(1,\ldots,|k|-1)$

(c) sample $|k_1|$ species among the $|k|$ species in group k and allocate them to the first subgroup k_1. The others are allocated to the second subgroup k_2. Let D denote the resulting allocation vector of the $|k|$ species between k_1 and k_2.

Let $\vec{C}^\star = (\vec{C}, k, |k_1|, D)$ denote the new classification that results from $\vec{C}$ through steps (a)–(c). Then, the conditional probability distribution of the new classification into $K + 1$ groups given the old one into K groups, $\pi^{\text{split}}_{\vec{C}^\star|\vec{C},K}$, is defined by:

$$\pi^{\text{split}}_{\vec{C}^\star|\vec{C},K}(\vec{C}^\star|\vec{C},K) = \Pr(\vec{C}^\star = (\vec{C},k,|k_1|,D)|\vec{C},K) = \frac{|k_1|!(|k|-|k_1|)!}{|k|!}\frac{1}{|k|-1}\frac{1}{\sum_{i=1}^{K}\mathbb{1}_{|i|>1}}\frac{1}{2}$$

• Death case: $K^* = K - 1$. The proposal for the latent class vector is obtained by merging two groups into a single one: randomly choose two groups among K and merge them into one group. Let k_1 and k_2 be the two selected groups and let $\vec{C}^\star = (\vec{C}, k_1, k_2)$ be the new classification that results from $\vec{C}$ by merging k_1 and k_2. Then, the conditional probability distribution of the new classification into $K - 1$ groups given the old one into K groups, $\pi^{\text{merge}}_{\vec{C}^\star|\vec{C},K}$, is defined by:

$$\pi^{\text{merge}}_{\vec{C}^\star|\vec{C},K}(\vec{C}^\star|\vec{C},K) = \Pr(\vec{C}^\star = (\vec{C},k_1,k_2)|\vec{C},K) = \frac{2!(K-2)!}{K!}\frac{1}{2}$$

2. Selection step. Given $\vec{C}$ and K, the vector of new parameters $\vec{\theta}^\star = (\vec{p}^\star, \vec{q}^\star, \vec{m}^\star, f^\star, \vec{d}^\star)$ is sampled from its marginal posterior distribution $\pi^{\underline{N}}_{\vec{\theta}|\vec{C},K}(\vec{\theta}|\vec{C},K,\underline{N})$. This marginal posterior distribution (not given here to save space) is known in an analytical form since multinomial/Dirichlet and Poisson/gamma distributions are conjugate distributions (Robert & Casella 2005).

The following equations give the expression of the Metropolis-Hasting ratio in the death case, for example. Let the current number of groups be K, and the new state K^* be $K - 1$. Let us assume that two groups k_1 and k_2 have been chosen and merged into a unique group k. Then,

$$\frac{\pi^{\text{split}}_{\vec{C}|\vec{C}^\star,K^\star}(\vec{C}|\vec{C}^\star,K^\star)}{\pi^{\text{merge}}_{\vec{C}^\star|\vec{C},K}(\vec{C}^\star|\vec{C},K)} = \frac{\binom{|k|}{|k_1|}\frac{1}{|k|-1}\frac{1}{\sum_{i=1}^{K}\mathbb{1}_{|i|>1}}}{\binom{K}{2}}$$

Moreover, $\frac{\pi^{\underline{N}}_{\vec{\theta}|\vec{C},K}(\vec{\theta}|\vec{C},K,\underline{N})}{\pi^{\underline{N}}_{\vec{\theta}|\vec{C},K}(\vec{\theta}^\star|\vec{C}^\star,K^\star,\underline{N})}$ is the ratio of marginal posterior distributions of $\vec{\theta}$ and is equal to

$$\frac{\pi^{\underline{N}_k}_{\theta}(\theta_k|\underline{N}_k)}{\pi^{\underline{N}_{k_1}}_{\theta}(\theta_{k_1}|\underline{N}_{k_1})\pi^{\underline{N}_{k_2}}_{\theta}(\theta_{k_2}|\underline{N}_{k_2})}$$

where $\underline{N}_k$ is the set of observations belonging to all species classified in group k. $\pi^{\underline{N}_k}_{\theta}(\theta|\underline{N}_k)$ is broken down as follows:

$$\pi^{\underline{N}_k}_{\theta}(\theta|\underline{N}_k) = \prod_{l}^{L}\pi^{\underline{N}_k}_{pqm|l,k}(p_l,q_l,m_l|\underline{N}_k)\;\pi^{\underline{N}_k}_{\vec{d}|k}(\vec{d}|\underline{N}_k)\;\pi^{\underline{N}_k}_{f|k}(f|\underline{N}_k)$$

where

$$\pi^{\underline{N}_k}_{pqm|l,k} \equiv \mathcal{D}(1+n_{llk}, 1+n_{l(l+1)k}, 1+n_{l\dagger k})$$

where n_{llk}, $n_{l(l+1)k}$ and $n_{l\dagger k}$ are the number of individuals in group k that respectively stay in class l, move from class l to $l + 1$ or die;

$$\pi^{\underline{N}_k}_{\vec{d}|k} \equiv \mathcal{D}(1+n_{1k},\ldots,1+n_{Lk})$$

where n_{lk} is the number of individuals of group k in class l at initial time t; and finally,

$$\pi^{\underline{N}_k}_{f|k} \equiv \mathcal{G}\left(0.01+n_{01k}, \frac{1}{n_k+1}\right)$$

where n_k is the total number of individuals in group k at initial time t and n_{01k} is the number of recruits in group k. Given this, the calculation of prior distribution as well as likelihood ratios is straightforward. As the matrix population model parameters are sampled from their posterior distributions, the canonical reversible transition function is the identity function. Hence, its Jacobian is equal to one and does not appear in the Metropolis-Hasting ratios.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Data S1. R scripts for the Bayesian inference algorithm.

Data S2. Parameters of the mixture matrix models with five tree species groups at Paracou, French Guiana.

Mixture of inhomogeneous matrix models for species-rich ecosystems

Environmetrics 2015

Research Article **Environmetrics**

Received: 22 May 2014, Revised: 12 October 2014, Accepted: 14 October 2014, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2320

Mixture of inhomogeneous matrix models for species-rich ecosystems

Frédéric Mortier[a]*, Dakis-Yaoba Ouédraogo[a], Florian Claeys[a,b,c], Mahlet G. Tadesse[d], Guillaume Cornu[a], Fidèle Baya[e], Fabrice Benedet[a], Vincent Freycon[a], Sylvie Gourlet-Fleury[a] and Nicolas Picard[a]

Understanding how environmental factors could impact population dynamics is of primary importance for species conservation. Matrix population models are widely used to predict population dynamics. However, in species-rich ecosystems with many rare species, the small population sizes hinder a good fit of species-specific models. In addition, classical matrix models do not take into account environmental variability. We propose a mixture of regression models with variable selection allowing the simultaneous clustering of species into groups according to vital rate information (recruitment, growth and mortality) and the identification of group-specific explicative environmental variables. We develop an inference method coupling the R packages `flexmix` and `glmnet`. We first highlight the effectiveness of the method on simulated datasets. Next, we apply it to data from a tropical rain forest in the Central African Republic. We demonstrate the accuracy of the inhomogeneous mixture matrix model in successfully reproducing stand dynamics and classifying tree species into well-differentiated groups with clear ecological interpretations. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: mixture models; lasso selection; species-rich ecosystems; usher models

1. INTRODUCTION

Understanding how environmental factors could impact population dynamics is of primary importance for animal and plant species conservation. Mathematical and statistical models are required to understand and predict these dynamics (Fieberg and Ellner, 2001; Demyanov *et al.*, 2006). Habitat models (Pearson *et al.*, 2002; Hargrove and Hoffman, 2004; García-López and Allué, 2011) use the spatial distribution of climate variables to predict the spatial range of species. These models are static in space and time and are conceptually unable to deal with situations where species are not in equilibrium with their environments (Stankowski and Parker, 2010). Ecophysiology-based dynamic global vegetation models (e.g., Scheiter and Higgins, 2009) precisely describe the biological processes that underlie growth, mortality and recruitment but require a huge amount of information. In species-rich ecosystems, limited information is available for each species. It is thus intractable to characterize different species with these models; instead, a plant functional type assumed to be representative of several species is modelled. As a consequence, these methods are more useful to predict biome changes at a continental scale than forest changes at a regional scale. Gap models (Solomon, 1986; Pastor and Post, 1988; Prentice *et al.*, 1993; Shao, 1996; Talkkari *et al.*, 1999), while using a simplified description of biological processes when compared with process-based models, still suffer from the same information limitation and are hardly used for species-rich forest ecosystems (Shugart and West, 1980).

Matrix population models, on the other hand, have been widely used to investigate the dynamics of age-, stage- or size-structured populations (Caswell, 2001; Stott *et al.*, 2010). They provide a simple way of integrating vital rate information such as birth, recruitment, growth or ageing and mortality (Crone *et al.*, 2011; Liang, 2010). In forest ecology and forest management, matrix models have been used to study natural successions, biodiversity dynamics and the impact of natural disturbances. They have also been used to evaluate economic outcomes and ecological impacts and to optimize management strategies (Buongiorno and Gilless, 2003).

Another challenge with species-rich ecosystems, such as tropical rain forests, tropical marine fish or coral reefs, is their high diversity, which implies that the sample size for most species is limited. The small sample size hinders development of species-specific models.

* *Correspondence to: F. Mortier, UPR Bsef, CIRAD, TA C-105/D, Campus International de Baillarguet, Montpellier, 34398, France. E-mail: fmortier@cirad.fr*

a *UPR Bsef, CIRAD, Montpellier, France*

b *AgroParisTech, Paris, France*

c *UMR Lef, AgroParisTech-INRA, Nancy, France*

d *Department of Mathematics and Statistics, Georgetown University, Washington, DC, U.S.A.*

e *Ministère des Eaux, Forêts, Chasse et Pêche, Bangui, Central African Republic*

To address this problem, modellers usually cluster species into groups using a variety of methods (Swaine and Whitmore, 1988; Steneck and Dethier, 1994; Favrichon, 1994; Bellwood and Wainwright, 2001; Gitay and Noble, 1997). Mixture models that cluster based on similar species responses rather than similar species traits have been proposed in the framework of generalized linear models (GLM) (Dunstan *et al.*, 2011; Dunstan *et al.*, 2013; Hui *et al.*, 2013; Ouédraogo *et al.*, 2013) and more recently in the context of homogeneous matrix population models (Mortier *et al.*, 2013).

In this paper, we propose a new class of mixture of inhomogeneous matrix population models that allows the simultaneous clustering of species based on vital rate processes (recruitment, growth and mortality) and selection of group-specific explicative environmental variables. The novelty of this method is that it provides the flexibility of selecting cluster-specific covariates in the context of multivariate GLM. It generalizes previous work for variable selection in multivariate Gaussian regression models (Brown *et al.*, 1998; Monni and Tadesse, 2009; Ouédraogo *et al.*, 2013) or in univariate GLM (Gupta and Ibrahim, 2007; Khalili and Chen, 2007; Städler *et al.*, 2010).

Section 2.2 is dedicated to the formulation of adaptive lasso regression mixture models and the associated expectation–maximization (EM) algorithm. Section 3 describes the simulation studies and a real dataset from the M'Baïki tropical rain forest in the Central African Republic, and Section 4 presents the corresponding results. The simulations demonstrate the effectiveness of the proposed method under various scenarios, while the real dataset highlights the performance of the mixture of inhomogeneous matrix models to predict stand characteristics of species-rich ecosystems in contrasted environmental conditions.

2. MODELS

2.1. Usher model

We first focus on a specific population labelled s and discuss the general setting that considers the whole stand in Section 2.2. The Usher matrix model applies to size-structured populations (Usher, 1966, 1969). It is based on the description of the change in the population size by a vector $N_s(t)$ containing the number of individuals in I ordered size classes at a discrete time t : $\mathbf{N}_s(t) = (N_{si}(t))_{i=1,\dots I}$, where $N_{si}(t)$ is the number of trees in the diameter class i at time t. The transitions between t and $t+1$ follow the Usher assumption that a tree can either stay in the same class, move up to the next class or die (moving backwards or moving up by more than one class are not allowed). The temporal change between times t and $t+1$ is defined by the recurrence relation

$$\mathbf{N}_s(t+1) = \mathbf{A}_s(t)\,\mathbf{N}_s(t) + \mathbf{R}_s(t) \tag{1}$$

where $\mathbf{A}_s(t)$ is the Usher $I \times I$ transition matrix for population s,

$$\mathbf{A}_s(t) = \begin{pmatrix} p_{s1}(t) & 0 & \dots & 0 \\ q_{s2}(t) & p_{s2}(t) & & 0 \\ & \ddots & \ddots & \\ 0 & & q_{sI}(t) & p_{sI}(t) \end{pmatrix} \tag{2}$$

and $\mathbf{R}_s(t)$ is the I-vector of recruitment for population s:

$$\mathbf{R}_s(t) = \begin{pmatrix} r_s(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3}$$

The transition parameters consist of: the stasis rate, $p_{si}(t)$, which corresponds to the probability of a tree in diameter class i at time t to stay alive and remain in the same diameter class at time $t+1$; the upgrowth rate, $q_{s,i+1}(t)$, which corresponds to the probability of a tree in diameter class i at time t to stay alive and to move up to diameter class $i+1$ at time $t+1$; and the recruitment flow, $r_s(t)$, which corresponds to the number of newly recruited trees in the first diameter class at time t. The transition parameters can be reparameterized as

$$\begin{aligned} q_{s,i+1}(t) &= q^{\bullet}_{s,i+1}(t) \times (1 - m_{si}(t)) \\ p_{si}(t) &= 1 - m_{si}(t) - q_{s,i+1}(t) \end{aligned} \tag{4}$$

where $q_{s,I+1}(t) = 0$ and $q^{\bullet}_{s,i+1}(t)$ is the conditional probability for a tree in diameter class i at time t to move up to diameter class $i+1$ given that it stays alive, and $m_{si}(t)$ is the probability for a tree in diameter class i to die between times t and $t+1$. Recruitment is assumed additive rather than proportional to the number of trees in each diameter class (Buongiorno and Michie, 1980). This means that the recruitment flow does not follow from the population alone but also involves an external inflow from the surrounding community. This additive recruitment is suited to the M'Baïki experimental case, where the observed plots are a sample of the whole forest (Caswell, 2001). A particular aspect of this matrix model is that the transition matrix $\mathbf{A}_s(t)$ and the recruitment vector $R_s(t)$ have explicit time dependence introduced through the linear associations of the demographic processes with time-varying environmental covariates. This contrasts with standard matrix models that are stationary.

2.1.1. *Predicting growth*

The upgrowth transition rate $q^{\bullet}_{s,i+1}(t)$ is computed from $a_{si}(t)$, defined as the 'typical' diameter at breast height (dbh) growth rate of a tree in class i at time t. Let u_i and u_{i+1} be the boundaries of class i and let τ be the time step of the matrix model. All trees with dbh ranging from $u_{i+1} - a_{si}(t)\,\tau$ to u_{i+1} will grow up to the next class, whereas trees with a diameter ranging from u_i to $u_{i+1} - a_{si}(t)\,\tau$ will remain in the same class. The proportion of trees that grow up to the next diameter class can thus be computed as

$$q^{\bullet}_{s,i+1} = \frac{a_{si}(t)\,\tau}{d_i} \tag{5}$$

where $d_i = u_{i+1} - u_i$ is the width of diameter class i. The typical dbh growth rate $a_{si}(t)$ can be estimated using growth data from class i only or using a regression model that relates growth and size over the entire size range (Rogers-Bennett and Rogers, 2006). The advantages and limitations of each estimator have been discussed elsewhere (Picard *et al.*, 2008). Here, we use the regression approach and predict the typical dbh growth rate as

$$a_{si}(t) = X^{G}_{si}(t)\beta_s \tag{6}$$

where the β_s's are population-specific coefficients to be estimated from the data and $X^{G}_{si}(t)$ are a set of known time-varying environmental covariates associated to the growth process.

2.1.2. *Predicting mortality*

The probability $m_{si}(t)$ that a tree in diameter class i dies between times $t-1$ and t is computed as

$$m_{si}(t) = \text{logit}^{-1}\left[X^{M}_{si}(t)\gamma_s\right] \times (\tau/\Upsilon) \tag{7}$$

where $\text{logit}^{-1}(x) = (1 + \exp(-x))^{-1}$ is the inverse logit function, the γ_s's are population-specific coefficients to be estimated from the data, $X^{M}_{si}(t)$ are a set of known time-varying environmental covariates associated to the mortality process, and Υ is the time step for death observations. The ratio τ/Υ must ensure that $m_{si}(t) < 1$, which in practice is satisfied even when τ is 10-fold Υ because of the very small value of the inverse logit term.

2.1.3. *Predicting recruitment*

The number of recruits $r_s(t)$ at time t in the first diameter class is computed as

$$r_s(t) = \exp\left[X^{R}_{s}(t)\alpha_s\right] \times (\tau/\Upsilon) \tag{8}$$

where the α_s's are population-specific coefficients to be estimated from the data, $X^{R}_{s}(t)$ are a set of known time-varying environmental covariates associated to the recruitment process, and Υ is the time step for recruitment observations.

2.2. Mixture of regression models and variable selection

So far, we have considered a single population. We now consider the whole stand, with as many populations as there are species. Because there are a lot of species with very few individuals, the parameters α_s, β_s and γ_s cannot be estimated for all the species of the stand. Thus, we aim to group species based on their common behaviour (growth, mortality or recruitment) as well as their similar association patterns with environmental factors. Species in the same group will share the same estimated parameters.

Species clustering is defined separately for growth, recruitment and mortality processes, and the clustered responses are related to the predictors defined in Equations (6)–(8). We develop a unified method to simultaneously (i) classify species according to their response to the predictors, (ii) select the significant predictors and (iii) estimate the parameters α_s, β_s, and γ_s of Equations (6)-(8) for each species group. We use a finite mixture of GLM to classify species into groups and estimate the model parameters, and we incorporate an adaptive lasso penalty to select the predictors for each group (Städler *et al.*, 2010).

Let S be the number of species, T the number of measurement times, n_{st} the number of trees from species s measured at time t (where $s = 1, \ldots, S$ and $t = 1, \ldots, T$), and $n = \sum_{s=1}^{S}\sum_{t=1}^{T} n_{st}$ the total number of observations in the dataset. The time considered here is a chronological one used to model annual differences and does not correspond to tree age. Let $\mathbf{Y}$ be the random vector of observations associated with either growth increments or death events. We assume that the growth rate for a tree from species s in dbh class i (conditionally on the tree staying alive) follows a Gaussian distribution with expectation equal to $a_{si}(t)$ and variance σ^2_s and that the death event is distributed as a Bernoulli random variable with probability $m_{si}(t)$. Using mixture models to group species with similar characteristics, the log-likelihoods of the growth and mortality processes for the n observations are computed as

$$\ell_n(\boldsymbol{\psi}|\mathbf{Y}) = \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} \log\left[\sum_{k=1}^{K} \pi_k\, f(Y_{stj}|\mathbf{X}, \boldsymbol{\psi}_k)\right] \tag{9}$$

where K is the number of species groups, π_k is the mixing proportion of group k, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K)$ with $\boldsymbol{\psi}_k$ the model parameters for group k, and $\mathbf{X}$ is the design matrix of explanatory variables. For the growth model, f is the Gaussian density function, $Y_{stj} = \Delta D_{stj}/\Upsilon$, where ΔD_{stj} is the diameter increment between times t and $t + \Upsilon$ for the j-th tree from species s and Υ is the time step between successive observations, and $\boldsymbol{\psi}_k = (\beta_k, \sigma_k)$. For the mortality model, f is the Bernoulli probability mass function, $Y_{stj} = M_{stj}$, where M_{stj} is a binary indicator of whether the j-th tree from species s died between times $t-1$ and t, and $\boldsymbol{\psi}_k = \gamma_k$.

The log-likelihood for the recruitment process is given by

$$\ell_n(\boldsymbol{\psi}|\mathbf{Y}) = \sum_{s=1}^{S}\sum_{t=1}^{T} \log\left[\sum_{k=1}^{K} \pi_k \, f(Y_{st}|\mathbf{X}, \alpha_k)\right] \tag{10}$$

where f is the probability mass function associated to the Poisson distribution with expected value $\exp(\mathbf{X}\alpha_k)$, $Y_{st} = R_{st}$ is the observed number of recruited trees for species s at time t, and $\boldsymbol{\psi}_k = \alpha_k$. It should be noted that the Poisson distribution is restrictive because of its assumption of equal expectation and variance, which is often not satisfied for ecological count data (Flores *et al.*, 2009). The negative binomial distribution can be a solution but may not be sufficient to accommodate the large number of zeros often recorded for recruitment processes. An alternative would be to use zero-inflated distributions (Poisson or negative binomial).

The relevant covariates associated to the different processes may vary from one group to another. We propose using the adaptive lasso approach to select the group-specific covariates (Zou, 2006; Städler *et al.*, 2010). The estimator $\hat{\boldsymbol{\psi}}$ for the model parameters $\boldsymbol{\psi}$ then corresponds to the maximum of a penalized log-likelihood:

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \{\ell_n(\boldsymbol{\psi}|\mathbf{Y}) - \mathcal{P}_n(\boldsymbol{\psi})\}$$

where $\mathcal{P}_n$ is the adaptive lasso penalty:

$$\mathcal{P}_n(\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \eta_{nk} \sum_{l=1}^{L} \frac{|\psi_{kl}|}{|\hat{\psi}_{kl}|} \tag{11}$$

with ψ_{kl} the lth element of $\boldsymbol{\psi}_k$, $|\hat{\psi}_{kl}|$ the maximum likelihood estimator of ψ_{kl}, and η_{nk} a parameter selected using cross-validation.

2.2.1. *Expectation–maximization algorithm*

Because of the sum within the log in Equations (9) and (10), the penalized log-likelihood cannot be maximized analytically but can be numerically maximized using the EM algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 2008). The EM algorithm is an iterative procedure that alternates between two steps, the E (or expectation) step and the M (or maximization) step. It starts with a random assignment of the species to the K groups. This gives the initial values $w_{stjk}^{(0)}$ of the posterior probability that the j-th tree from species s at time t belongs to species group k : $w_{stjk}^{(0)} = 1$ if species s is initially assigned to group k, and 0 otherwise.

In the E-step, the posterior probability that the j-th tree from species s at time t belongs to species group k is computed as

$$w_{stjk}^{(m)} = \frac{\pi_k^{(m)} \prod_{t'=1}^{T} \prod_{j'=1}^{n_{st'}} f\left(Y_{st'j'}|\mathbf{X}, \boldsymbol{\psi}_k^{(m)}\right)}{\sum_{l=1}^{K} \pi_l^{(m)} \prod_{t'=1}^{T} \prod_{j'=1}^{n_{st'}} f\left(Y_{st'j'}|\mathbf{X}, \boldsymbol{\psi}_l^{(m)}\right)} \tag{12}$$

where the superscript m is the iteration index of the algorithm. An important point to notice is that $w_{stjk}^{(m)}$ does not depend on t and j. This is peculiar to situations with replicate measurements for the clustered unit and ensures that when a species is assigned to a group, all its conspecifics are also assigned to the same group. In other words, posterior group probabilities are computed at the species level rather than at the individual tree level. We adopt the approximation used in Khalili and Chen (2007) to update the mixing proportions as

$$\pi_k^{(m+1)} = \frac{1}{n} \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} w_{stjk}^{(m)}$$

An improved update of the mixing proportions is provided in Städler *et al.* (2010).

In the M-step, the penalized log-likelihood is maximized for each component separately using the posterior probabilities of the observations as weights. This gives estimates for component k's parameters at the m-th iteration of the algorithm as

1. For the growth process

$$\hat{\beta}_k^{(m)} = \arg\max_{\beta_k} \left\{\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} w_{stjk}^{(m-1)} \log f\left(\Delta D_{stj}/\Upsilon | X_{kj}^{G} \beta_k, \sigma_k^2\right) - \pi_k^{(m-1)} \eta_{nk} \frac{|\beta_k|}{|\hat{\beta}_k|}\right\} \tag{13}$$

where f is the density of the Gaussian distribution.

2. For the death process

$$\hat{\gamma}_k^{(m)} = \arg\max_{\gamma_k} \left\{ \sum_{s=1}^{S} \sum_{t=1}^{T} \sum_{j=1}^{n_{st}} w_{stjk}^{(m-1)} \log f\left(M_{stj} | X_{kj}^{M} \gamma_k\right) - \pi_k^{(m-1)} \eta_{nk} \frac{|\gamma_k|}{|\hat{\gamma}_k|} \right\} \tag{14}$$

where f is the probability mass function associated to the Bernoulli distribution.

3. For the recruitment process

$$\hat{\alpha}_k^{(m)} = \arg\max_{\alpha_k} \left\{ \sum_{s=1}^{S} \sum_{t=1}^{T} w_{stk}^{(m-1)} \log f\left(R_{st} | X_k^{R} \alpha_k\right) - \pi_k^{(m-1)} \eta_{nk} \frac{|\alpha_k|}{|\hat{\alpha}_k|} \right\} \tag{15}$$

where f is the probability mass function associated to the Poisson distribution.

2.2.2. *Number of components and species allocations*

The model fitting described in the previous paragraphs supposes that the number of groups K is known. In order to estimate it, we fit the finite mixture of GLM for $K = 1, 2, 3, \ldots$, and we select the value of K that minimizes an information criterion. Different criteria have been used, such as the Akaike information criterion (Akaike, 1974), the Bayesian information criterion (Schwarz, 1978), or the integrated completed likelihood criterion (ICL) (Biernacki *et al.*, 2000). We adopt the ICL, which has been specifically developed for mixture models and takes into account the quality of the classification. The ICL penalization is given by

$$\mathcal{P}_{\text{ICL}} = \nu_K log(n) + 2 \sum_{s=1}^{S} n_s \sum_{k=1}^{K} w_{sk} \log(w_{sk})$$

where the first term corresponds to the Bayesian information criterion penalization with ν_K equal to the number of free parameters in the model with K components, $n_s = \sum_{t=1}^{T} n_{st}$ is the number of tree observations for species s, and w_{sk} is the estimated posterior probability that species s belongs to group k (Equation 12). The maximum *a posteriori* estimate is then used to determine each species' allocation.

2.3. Mixture of inhomogeneous matrix models

The mixture of GLM gives K_g species groups for growth, K_r for recruitment and K_m for mortality. Crossing these classifications gives $K_g \times K_r \times K_m$ combinations of groups. These combinations are named $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$, with $1 \leqslant x \leqslant K_g, 1 \leqslant y \leqslant K_r$, and $1 \leqslant z \leqslant K_m$. Because of the additive recruitment, each of the K_r recruitment groups contributes to several combinations of groups. Therefore, the number of recruits $r_y(t)$ for recruitment group y must be distributed between the combinations $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$. The estimated number of recruits for the combination $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$ is computed as $\rho_{xyz}\, r_y(t)$, where $\rho_{xyz} = N_{xyz} / \sum_{x'} \sum_{z'} N_{x'yz'}$ is the ratio of the total number of alive trees in combination $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$, N_{xyz}, over the total number of alive trees in recruitment group y, such that $\sum_x \sum_z \rho_{xyz} = 1$ for all y.

Each species exclusively belongs to one combination of groups. Because the parameters of the growth, mortality and recruitment models are estimated for each group, the look-up table assigning each species to growth group x, recruitment group y and mortality group z defines a matrix population model for it. Therefore, the combinations of groups define what we call *the mixture of inhomogeneous matrix models*.

3. APPLICATION

3.1. Simulations

We simulated mixture regression models with the true number of components set to three. We generated 30 species, and within each species, we sampled the number of trees from a Poisson(30). Within each tree in a given species, the number of repeated measures was sampled from a Poisson(15). To evaluate the effect of ignoring the time dependence in our model, we considered a first-order autoregressive correlation structure (AR1(ρ)) with varying correlation parameters $\rho = (0, 0.1, 0.3, 0.5, 0.7, 0.9)$; this autoregressive dependence was applied on the residuals for the Gaussian case and on the linear predictors for the Bernoulli and Poisson cases. The species were randomly assigned to the three groups with mixing proportions set to $\pi = (0.60, 0.25, 0.15)$. A total of five covariates were generated from a multivariate normal distribution with mean 0 and an AR1(0.7) covariance matrix. We also considered a scenario where the design matrix $\mathbf{X}$ has dependence structure across covariates with correlation of 0.5 in addition to the temporal AR1(0.7) correlation for repeated measures within the same covariate. The parameters associated to each covariate had a 0.5 probability of being zero, and the nonzero parameters were simulated as described in Table 1. We generated 50 datasets. For each simulation, a K-component mixture model was fit three times with different starting points for $K = 1, \ldots, 7$. We retained the fit and the K value that yield the lowest ICL. The computations were performed using the R software (R Core Team, 2014) by integrating functionalities of the `flexmix` (Leisch, 2004; Grün and Leisch 2007, 2008) and the `glmnet` (Friedman *et al.*, 2010) packages (see the Supporting information for the complete R code). For the algorithm to converge, it is necessary to use the same cross-validation partitioning across the EM iterations, that is, the subsamples for cross-validation must be defined at the beginning using the `foldid` option in the function `FLXMRglmnet` (see documentation in `glmnet`).

Table 1. Parameters used for simulations

Distribution	Intercept	Covariates coefficient	Variance
Gaussian	$\{-1.5, 0, 1.5\}$	$\mathcal{U}[-2, -1, 1, 2]$	1
Bernoulli	$\{-1.5, 0, 1.5\}$	$\mathcal{U}[-2, -1, 1, 2]$	—
Poisson	$\{-1, 0, 1\}$	$\mathcal{U}[-1, 1]$	—

Intercepts are fixed, one for each group, along a gradient of values set to -1.5, 0, 1.5 in the Gaussian and Bernoulli cases and equal to -1, 0, 1 in the Poisson case. The nonzero coefficients associated to the relevant covariates are randomly drawn from a discrete uniform distribution $\mathcal{U}$ in the set of values given between brackets.

3.2. The M'Baïki forest case study

3.2.1. The experimental site

We applied the method to the M'Baïki species-rich tropical rainforest ecosystem. The M'Baïki experimental site (3°54'N, 17°56'E) was established in a lowland semi-deciduous tropical rain forest of the Central African Republic. The average annual rainfall for the period 1981 – 2008 is 1739 mm with a 4-month dry season and an annual average monthly temperature of 24.9°C (Ouédraogo *et al.*, 2013). The M'Baïki experimental site consists of 10 permanent sample plots, each of 4 ha (200 m × 200 m), established in two forests less than 10 km apart (Figure 1). Two blocks of three plots each were established in the Boukoko forest and one block of four plots in the La Lolé forest (Bedel *et al.*, 1998). These permanent sample plots have been inventoried every year since 1982 (except in 1997, 1999 and 2001): all trees $\geqslant$ 10 cm dbh have been individually marked and spatially located and have been measured yearly for dbh. All species present have been identified, and dead trees and newly recruited trees with dbh $\geqslant$ 10 cm have been surveyed. The type of soil in all plot, except one, has been mapped.

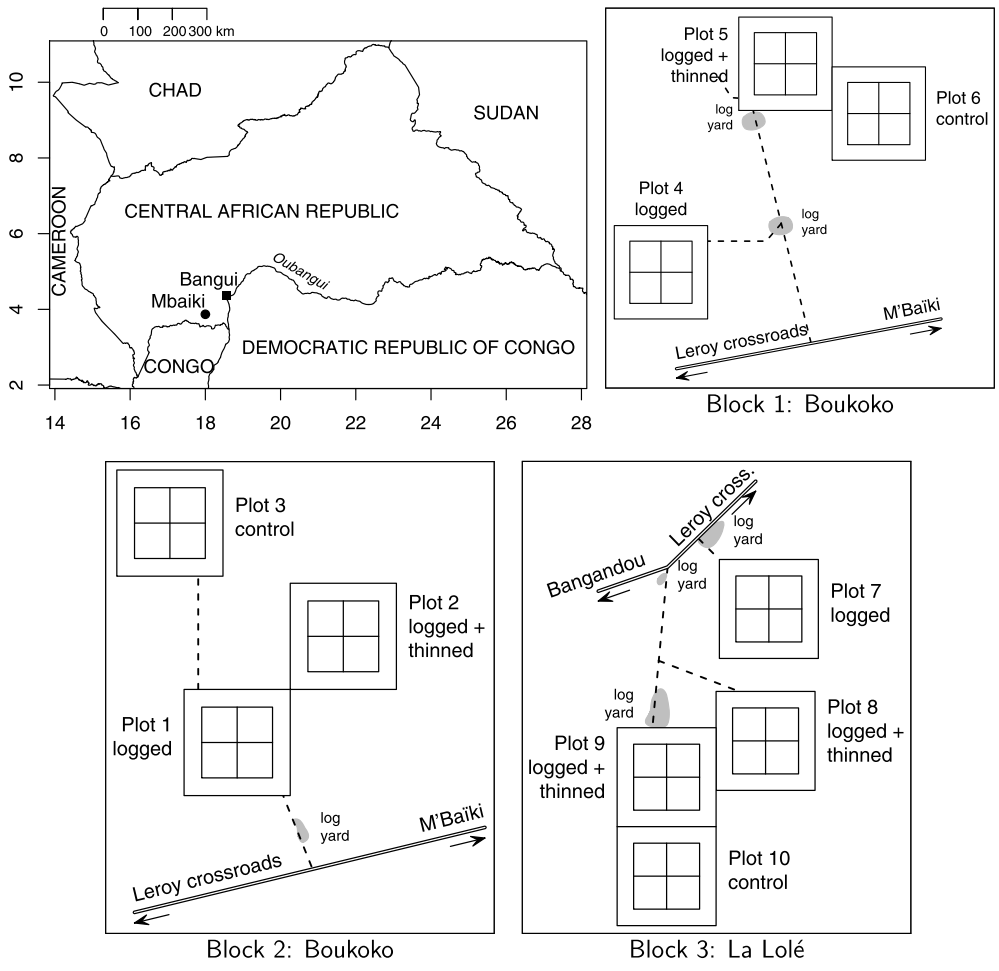


Figure 1. The M'Baïki forest experimental plots in the Central African Republic

Seven of the 10 plots across the three blocks were selectively logged between the 1984 and 1985 inventories. Three plots, one from each block, were left as controls. Logging consisted in harvesting trees with dbh $\geqslant$ 80 cm if belonging to one of 16 commercial species. Four of the seven plots logged (one from each of the Boukoko blocks and two from the La Lolé block) were thinned 2 years after logging to increase light penetration. Thinning consisted in poison girdling all nontimber trees with dbh $\geqslant$ 50 cm. This process was completed by cutting all lianas in the entire plot. The M'Baïki experimental site thus provides a perfect setting to observe the demographic processes across a wide range of disturbances, from undisturbed forests (unlogged plots) to highly disturbed forests (logged + thinned plots). Between 1982 and 2012, more than 37 000 trees from 230 genera have been monitored at this site. For this study, years for which complete data on the demographic processes and environmental variables are available were considered for analysis, resulting in $T = 18$.

3.2.2. *Growth, mortality and recruitment quantification*

The observations use an annual time step ($\Upsilon = 1$). To quantify the annual tree growth process, we calculated the annual tree diameter increments using only measurements from living trees that exhibit no trunk anomalies between two successive years. To further eliminate measurement errors, we only kept diameter increments between -0.4 cm (corresponding to stem shrinkage during dry seasons) (Baker *et al.*, 2002) and 4.456 cm, the 99th percentile of observed diameter increments of the fastest growing species *Musanga cecropioides* (Ouédraogo *et al.*, 2013).

The data were split into a training and a validation sets. The training dataset is taken to be Block 2 from the Boukoko forest and consists of three plots with the three different treatments (Figure 1). This block contains 197 species out of the 230 identified across all the M'Baïki plots and has data on 80 510 growth observations, 118 133 mortality observations and 42 816 recruitment observations. It is used to fit the growth, mortality and recruitment processes. The validation dataset consists of the other block in Boukoko and the block in La Lolé. It is used to evaluate the prediction quality of the mixture of inhomogeneous matrix models for plots sampled in contrasted environmental conditions.

To limit the discretization bias that may result from matrix modelling (Shimatani *et al.*, 2007; Picard *et al.*, 2010; Zuidema *et al.*, 2010), we use very thin dbh classes with a width of $d = 1$ cm. The time interval of the model has to be adjusted to the class width to meet the Usher assumption. This is achieved with a short time step of $\tau = 0.1$ year.

3.2.3. *Environmental covariates*

Five environmental variables and two variables describing the tree development stage were considered as potential covariates for the growth and mortality processes. The latter variables are the dbh and log-dbh (Di in cm and log-Di), which are commonly included in the model simultaneously to deal with the nonlinear association between dbh and growth (or mortality) (Zeide, 1993; Weiskittel *et al.*, 2011). The five environmental variables include two plot-level variables assessing competition for resources and three climate variables (see (Ouédraogo *et al.*, 2013) for details). The two competition indices are stand basal area (m^2 per hectares, BAst) and stand density (number of trees per hectares, Dst), which are computed on 1-ha subplots (100 m $\times$ 100 m) obtained as a subdivision of the initial 4-ha plots into four squares. This spatial unit was used because the environment is more homogeneous at this scale. The three climate variables are drought indices: the length of the dry season (number of months with rainfall $<$ 100 mm, LDS), the average rainfall during the dry season (RDS in millimetre) and the annual average soil water content (MSW in millimetre) (Ouédraogo *et al.*, 2013). For the recruitment process, potential predictors were restricted to BAst, Dst, LDS and RDS.

3.2.4. *Adjustment of the method to the M'Baïki forest*

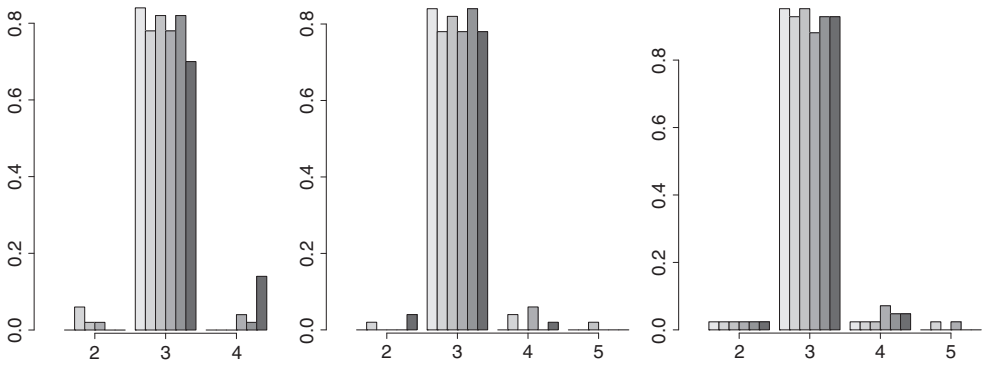
The models were fit for each process using $K = 1, \ldots, 10$ groups. This was repeated 10 times with different initial random points for each K, and the fit with smallest ICL was chosen. The group structures for the growth and mortality processes were successfully identified. However, because of the large number of zeros in the recruitment, the mixture model did not work as well. We therefore made some adjustments to adapt the inference for this process. We assumed that the species groups identified for the growth process are nested within the groups of the recruitment process. This assumption is supported by the well-established positive correlation between species-specific recruitment rates and growth rates in disturbed forests, which is a direct consequence of the recruitment design that requires passing a 10 cm dbh threshold (Gourlet-Fleury *et al.*, 2005). Therefore, once we identified the growth groups, the recruitment groups were obtained by fitting a mixture of Poisson regression models to the number of recruits of the growth groups, instead of the number of recruits of the species.

A second adjustment to the general framework presented earlier was made to deal with species that could not be classified for various reasons, including situations in which the species were not available in the training data, environmental covariates were missing for the species, or the species had a single individual measurement. The strategy we adopted is presented in Section 4.

4. RESULTS

4.1. Simulations

The algorithm performs quite well even when the dependence across time is not taken into account. We are able to identify the correct number of underlying clusters for all the different processes with correlations as high as 0.9 between consecutive repeated measures (Figure 2). We use two matching indices, I_1 and I_2 (Mortier *et al.*, 2013), to assess the clustering performance and compare each species group allocation based on the maximum *a posteriori* estimate to the true group membership. These indices are based on the $K \times \hat{K}$ contingency table



(a) Gaussian (b) Poisson (c) Bernoulli

Figure 2. Distribution of the estimated number of groups based on 50 replications of a simulated dataset with three groups, when observations have either a (a) Gaussian, (b) Poisson or (c) Bernoulli distribution. We considered a first-order autoregressive correlation structure (AR1(ρ)) with varying correlation parameters from 0 (light grey) to 0.9 (dark grey)

$C = (C_{ij})$ with $i = 1, \ldots, K$ and $j = 1, \ldots, \hat{K}$ that cross-tabulates the species according to their true and estimated classifications:

$$I_1 = \frac{1}{S}\sum_{i=1}^{K} \max\left\{C_{i1}, \ldots, C_{i\hat{K}}\right\} \qquad I_2 = \frac{1}{S}\sum_{j=1}^{\hat{K}} \max\{C_{1j}, \ldots, C_{Kj}\}$$

These indices vary between $1/S$ and 1 with higher values corresponding to better classifications. For $\hat{K} = K$, we obtain 98% of the time $I_1 = I_2 = 1$. When considering $\hat{K} = K + 1$ (which occurred rarely), we obtain 93% of the time $I_1 = 1$ and the few instances where $I_1 < 1$ are due to a group being split into two subgroups (I_2 is always lower than one by construction).

The algorithm is also effective at selecting the component-specific relevant covariates for all the distribution types (Gaussian, Bernoulli or Poisson). For example, in the more complex scenario where the design matrix **X** has both temporal dependence and correlated covariates, we obtain the following results: in the Gaussian case, out of the 50 simulations, one false positive is included one time; in the Bernoulli case, one false positive is selected five times, and there is a single instance of two false negatives; in the Poisson case, one, three or four false positives are selected one time each, and there is a single instance of one false negative.

4.2. The M'Baïki forest case study

4.2.1. Species classification and ecological meaning

Six groups are identified for the growth process, labelled g_1 to g_6 in order of increasing maximum growth rate, which is used as a proxy for light requirement. These six groups are nested within four recruitment groups, $r_1, \ldots, r_4 : g_2$ and g_6 correspond to r_2, g_5 and g_4 match with r_1, g_3 with r_4 and group g_1 constitutes r_3. We also identify three mortality groups, labelled m_1 to m_3. The growth ordering does not parallel the mortality ordering, and no obvious relationship can be found between growth and mortality groups. The ICL curves as well as parameter estimates are presented in the Supporting information.

Crossing these classifications gives $6 \times 4 \times 3 = 72$ possible combinations of groups, of which only 15 are nonempty. Accordingly, the mixture of matrix models is composed of 15 transition matrices. The nonempty combinations of groups contain between a single species up to 24 species (with known regeneration guild, (Bénédet *et al.*, 2014)) and correspond to groupings that are biologically meaningful, especially in terms of regeneration guild (Table 2). Moreover, the clusters uncovered by the mixture of Usher matrix models group species according to both their maximum growth rate and their maximum diameter (95th percentile). When plotting species along these two axes, the combinations of groups are well separated (Figure 3). Because these two axes can be used to order species along a continuum of ecological strategies (Turner, 2001; Alder *et al.*, 2002), this provides evidence that the mixture of inhomogeneous Usher matrix models is able to cluster species in a way that is consistent with their autecology (Picard *et al.*, 2012).

4.2.2. Prediction results, correction factors and asymptotic state

Among the 230 tree species at M'Baïki, 12 were not considered for analysis for various reasons (missing covariates and lack of replicate measurements) and remained unclassified. Out of the 218 tree species retained for analysis, 21 are not present in the training set but are present in the validation dataset and are classified *a posteriori*. It is still necessary to account for the 12 unclassified species when computing the stand basal area (Bast(t)) and the stand density (Dst(t)) to avoid underestimating these two competition indices. Hence, correction factors c_B and c_D are applied to Bast(t) and Dst(t), respectively. Factor c_B is computed as the ratio of the total stand basal area in 1992 over the cumulated basal area of classified species in 1992: $c_B = 1.00259\,(\pm 0.00027)$. Factor c_D is computed as the ratio of the total number of trees in 1992 over the cumulated number of trees from species that were classified in 1992: $c_D = 1.000351\,(\pm 0.00011)$.

Table 2. Floristic characteristics of the combinations of growth, recruitment and mortality groups identified at M'Baïki: number of species in each combination (size), regeneration guild (guild), phenology and dominant species.

	Classification characteristics			
Group	Size	Guild	Phenology	Dominant species
$g_1r_3m_3$	4	SB	Ever	Garcinia smeathmannii
$g_1r_3m_1$	20	NPLD-SB	Dec	Canarium schweinfurthii
$g_2r_2m_1$	24	NPLD-SB	Dec	Entandrophragma candollei
$g_2r_2m_2$	3	SB	Ever	Cola altissima
$g_2r_2m_3$	4	SB	Ever	Afrostyrax lepidophyllus
$g_3r_4m_2$	3	SB	Dec	Monodora myristica
$g_3r_4m_1$	24	NPLD-SB	Dec	Entandrophragma utile
$g_3r_4m_3$	1	P	Ind	Zanthoxylum lemairei
$g_4r_1m_2$	1	NPLD	Ever	Pycnanthus angolensis
$g_4r_1m_1$	22	NPLD	Dec	Entandrophragma angolense
$g_5r_1m_2$	1	P	Ind	Dictyandra arborescens
$g_5r_1m_1$	21	NPLD	Dec	Lovoa trichilioides
$g_5r_1m_3$	3	NPLD	Dec	Entandrophragma cylindricum
$g_6r_2m_3$	2	P	Ever	Cleistopholis glauca
$g_6r_2m_1$	11	P	Dec	Terminalia superba

SB, shade bearer; NPLD, nonpioneer light demander; P, pioneer; Ever, evergreen; Dec, deciduous; and Ind, unknown phenology.
Regeneration guild is determined for each group based on two aspects: the guild of the species with the largest number of trees in the group and the guild that contains the most species in the group. In most cases, the two agree, but when they are different, we provide both (e.g., NPLD-SB). Dominant species means that this species has the highest number of trees in the group.

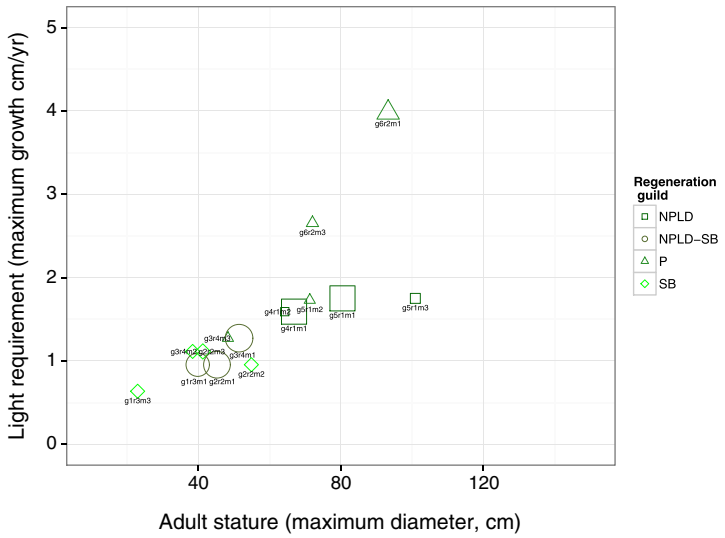


Figure 3. Projection of the species clustering obtained by the inhomogeneous mixture of Usher matrix models at M'Baïki on the two axes corresponding to the maximum diameter and the maximum growth rate. The labels $g_x r_y m_z$ correspond to the identified species groups. Each symbol corresponds to the dominant regeneration guild of each group. The size of the symbol is proportional to the number of species in the group

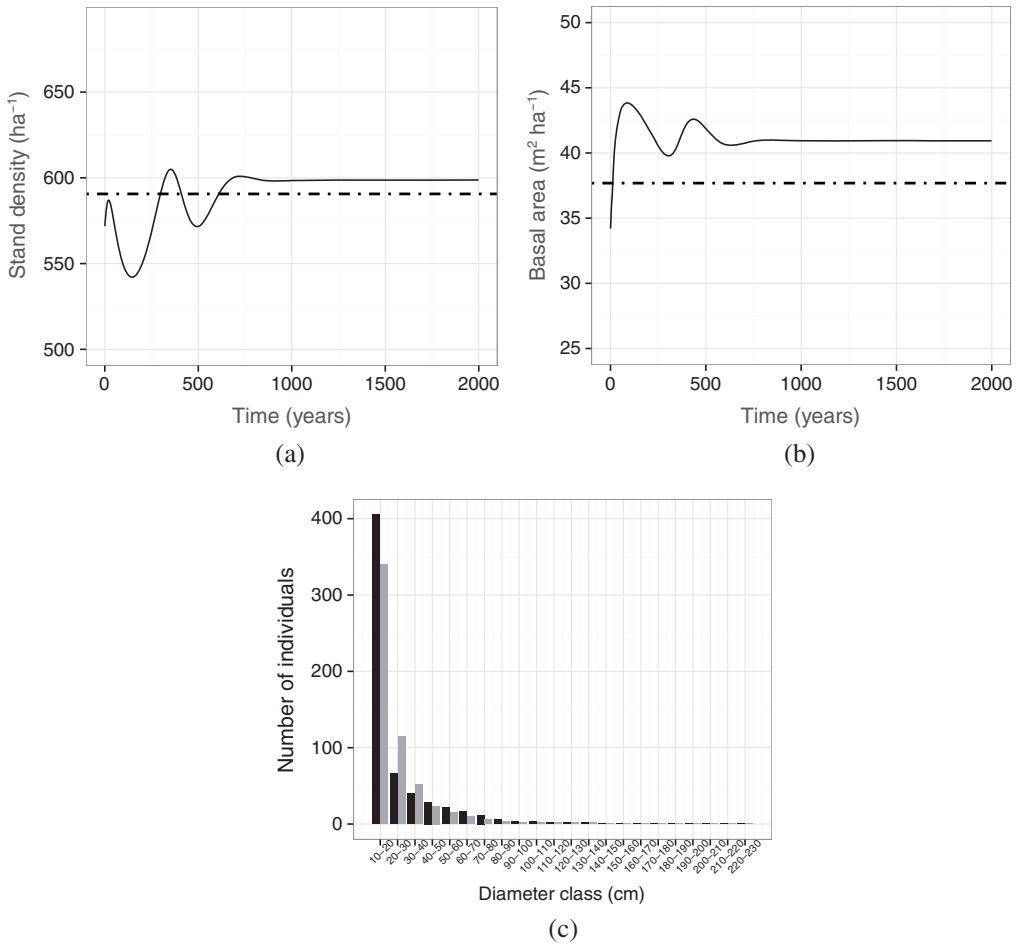


Figure 4. Density (number of tree per hectares), basal area (per hectares) and diametric structure. In (a) and (b), the solid lines correspond to the simulated forest and the dot-dashed lines to the observed stand in 2012 on the validation blocks. In (c), the black bars correspond to the simulated forest and the grey ones to the observed stand

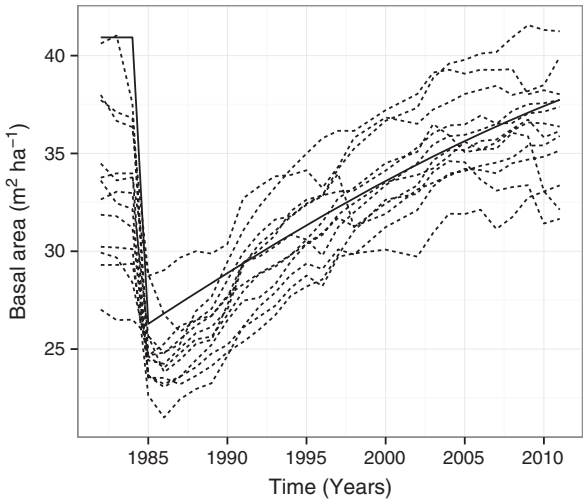


Figure 5. Dynamics of the basal area after logging (solid line: prediction; dashed lines: observations from 1982 to 2012 in the logged plots of the validation blocks)

Year 1992 is chosen because information for all processes and environmental variables is available from this time on. The two competition indices are then computed from the vector of the number of trees as $\mathrm{BAst}(t) = c_B \times \sum_s \mathbf{B}'\mathbf{N}_s(t)$ and $\mathrm{Dst}(t) = c_D \times \sum_s \mathbf{1}'\mathbf{N}_s(t)$, where $\mathrm{B} = \left(\frac{\pi}{4}\mathrm{D}_i^2\right)_{i=1\ldots I}$ is the vector of mean basal area for each diameter class, $\mathbf{1}$ is a vector of ones of length I and prime denotes the transpose operator.

The results of the simulated forest dynamics using the inhomogeneous matrix model over 2000 years starting with the observed forest stand in 1992 is shown in Figure 4 (see the Supporting information for the complete R code). The predicted asymptotic tree density, basal area and dbh structure match the observations of the validation data in 2012. In addition, the observed dbh distribution in 2012 at M'Baïki has an inverse-J shape that is typical of natural rain forests (Figure 4(c)). It could be fit by an exponential distribution with parameter 0.0724 (standard error 0.0047). In comparison, the predicted dbh distribution also presents an inverse-J shape and can be fit by an exponential distribution with parameter 0.0695.

We also compared the predicted dynamics following a 28-year wait after disturbance of the asymptotic state to the observed dynamics between 1982 and 2012 in the logged plots of the validation dataset (Figure 5). The simulated disturbance for the asymptotic state consisted of removing with probability 1/2 trees with dbh greater than 80 cm from the asymptotic dbh distribution. This corresponds to a perturbation of the same magnitude as the one realized in 1984 at M'Baïki in terms of lost basal area but performed on a wider range of species. The model successfully predicts the reconstitution rate of the basal area after disturbance (slope of dynamics): the predicted rate is 0.4329, while the observed rates in the logged plots of the validation data have a mean of 0.4517 and standard error 0.0929.

5. DISCUSSION

The proposed mixture of inhomogeneous matrix models is an original method that simultaneously fits matrix population models for species-rich ecosystems, clusters species into ecologically meaningful groups and selects relevant environmental covariates. As such, it is an integrated alternative to classical methods for building matrix population models, for classifying species or for selecting variables in regression models. The coupling of modern covariate selection methods and mixture model approaches that we have put forward in the mixture of inhomogeneous matrix models can be straightforwardly incorporated into any model where individual growth is regressed against size and environmental covariates. In particular, it could also be implemented in individual-based models (Dunstan *et al.*, 2011) or in integral projection models (Zuidema *et al.*, 2010).

Compared with other modelling approaches, the mixture of inhomogeneous matrix models combines the power of modern and technically complex statistical methods with the simplicity of matrix modelling. In this paper, we considered a few potential covariates, but the proposed method has the flexibility to handle a large number of covariates and select the relevant ones to model the dynamics and refine the predictions. For example, species-specific functional traits, such as the 99th percentile of diameter or wood density, as proposed by Hérault *et al.* (2011) could be included as potential covariates. For the front-end user, the model is as simple to use as any other matrix model. We thus expect the mixture of inhomogeneous matrix models to be useful in all application areas where matrix population models have been found to be useful decision tools, such as population viability analysis (Morris and Doak, 2002) or the management of wildlife population with harvest (Jensen, 1996), in particular when operating in a variable environment.

Taking into account environmental variability in matrix models is crucial to better understand and predict consequences of environmental variations on population dynamics. In the particular case of the M'Baïki tropical rain forest, we demonstrated the model's ability to reproduce the stand structure at equilibrium and the dynamics after disturbance. We showed, using simple exploitation rules, that the model could successfully reproduce post-logging dynamics over a 25-year period. Climate variables were also included in the environmental variables, thus paving the way for predicting the impact of climate change (Liang *et al.*, 2011), including the change in species composition or the interaction between disturbance and climate change, caused by the species differentiated responses to climate. The role of climate in forest dynamics at M'Baïki will be investigated in a future study.

Further work should be pursued to address some issues that were not taken into account in this paper. In particular, (i) explicitly modelling the time dependence between observations within the same tree, (ii) addressing the zero inflation in the recruitment process and (iii) investigating the impact of imbalanced class distributions on the results of the mixture models. For the first, mixed models offer a flexible method to handle longitudinal dependence (Bondell *et al.*, 2010; Schelldorfer *et al.*, 2014). Our method can be extended to accommodate this by considering mixtures of generalized linear mixed models with variable selection. However, this is computationally challenging and requires the development of efficient algorithms. For the second, zero-inflated distributions provide a general framework to overcome the presence of a large number of zeros (Flores *et al.*, 2009). However, the challenge of using zero-inflated models in the context of model-based clustering is the complexity of nesting two levels of mixtures: one corresponding to the mixture of a point mass at zero and a Poisson (or negative binomial) distribution and the other corresponding to the mixture of distributions used to identify groups of species. For the imbalanced class distribution issue, which may compromise the performance of clustering, sampling methods, such as random undersampling (Tseng and Wong, 2005), are commonly used to achieve a more balanced distribution. The integration of such sampling strategies with ensemble learning methods, such as bagging (Breiman, 1996) and boosting (Friedman, 2000), has been shown to improve the performance of imbalanced data classification/clustering (He and Garcia, 2009). However, the problem is more complicated in our context, where the clustering is performed at the species level and the imbalanced distribution occurs both at the level of the species and the varying number of trees within species.

Finally, we have fit the growth, mortality and recruitment models separately. This ensures an optimal fit for each dynamic component. However, because growth, mortality and recruitment are nonlinearly combined into the matrix model, this does not ensure an optimal fit at the matrix model level. Combining equations estimated separately may induce a prediction bias at the population level. Although scarcely documented in the scientific literature, this prediction bias is a well-known issue among forest modellers and occurs in different types of forest dynamic models. The problem is usually addressed by tuning *a posteriori* some coefficients (Favrichon, 1998). An alternative to deal

with this problem and a possible extension of our proposed model would be to formulate a unified approach that allows the fit of the three demographic processes simultaneously using an integrated population model (Abadi *et al.*, 2010). This can be achieved within a Bayesian hierarchical framework (Cressie *et al.*, 2009) by defining a first level that models the number of trees in a diameter class $\mathbf{N}_s(t)$ conditionally on the growth, mortality and recruitment processes and a second level that models these demographic processes using mixture models with variable selection similarly to the method we have proposed here.

Acknowledgements

This research was supported by the CoForChange project (http://www.coforchange.eu/) funded by the ERA-Net BiodivERsA with the national funders ANR (France) and NERC (UK), part of the 2008 BiodivERsA call for research proposals involving 16 European, African and international partners including a number of timber companies (see the list on the website, http://www.coforchange.eu/partners), and by the CoForTips project funded by the ERA-Net BiodivERsA with the national funders FWF (Austria), BelSPO (Belgium) and ANR (France), part of the 2011–2012 BiodivERsA call for research proposals (http://www.biodiversa.org/519). We would also like to thank the two anonymous referees for their constructive comments.

REFERENCES

Abadi F, Gimenez O, Arlettaz R, Schaub M. 2010. An assessment of integrated population models: bias, accuracy, and violation of the assumption of independence. *Ecology* **91**(1):7–14.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6):716–723.

Alder D, Oavika F, Sanchez M, Silva JNM, Van der Hout P, Wright HL. 2002. A comparison of species growth rates from four moist tropical forest regions using increment-size ordination. *International Forestry Review* **4**(3):196–205.

Baker TR, Affum-Baffoe K, Burslem D, Swaine MD. 2002. Phenological differences in tree water use and the timing of tropical forest inventories: conclusions from patterns of dry season diameter change. *Forest Ecology and Management* **171**(3):261–274.

Bedel F, Durrieu de Madron L, Dupuy B, Favrichon V, Maître H, Bar-Hen A, Narboni P. 1998. Dynamique de croissance dans des peuplements exploités et éclaircis de forêt dense africaine. le dispositif de m'baïki en république centrafricaine (1982-1995). *CIRAD Forêt, Montpellier Série FORAFRI, document* **1**:1–72.

Bellwood D, Wainwright P. 2001. Locomotion in labrid fishes: implications for habitat use and cross-shelf biogeography on the great barrier reef. *Coral Reefs* **20**(2):139–150.

Bénédet F, Vincke D, Fayolle A, Doucet F, Gourlet-Fleury S: Cofortraits, African plant traits information database. version 1.0. http://coforchange.cirad.fr/african_plant_trait, access to database can be granted upon request. [accessed on 24 September, 2013]

Biernacki C, Celeux G, Govaert G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7):719–725.

Bondell H, Krishna A, Ghosh S. 2010. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**(4):1069–1077.

Breiman L. 1996. Bagging predictors. *Machine Learning* **24**(2):123–140.

Brown P, Vannucci M, Fearn T. 1998. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60**(3):627–641.

Buongiorno J, Gilless J. 2003. *Decision Methods for Forest Resource Management*. Academic Press: Elsevier Science (USA).

Buongiorno J, Michie B. 1980. A matrix model of uneven-aged forest management. *Forest Science* **26**(3):609–625.

Caswell H. 2001. *Matrix Population Models, Construction, Analysis, and Interpretation* deuxième edition. Sinauer Associates, Inc. Publishers: Sunderland, Massachusetts.

Cressie N, Calder CA, Clark JS, Wikle CK. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications* **19**(3):553–570.

Crone E, Menges E, Ellis M, Bell T, Bierzychudek P, Ehrlén J, Kaye T, Knight T, Lesica P, Morris W, Oostermeijer G, Quintana-Ascencio P, Stanley A, Ticktin T, Valverde T, Williams J. 2011. How do plant ecologists use matrix population models? *Ecology Letters* **14**(1):1–8.

Dempster A, Laird N, Rubin D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39**(1):1–38.

Demyanov V, Wood S, Kedwards T. 2006. Improving ecological impact assessment by statistical data synthesis using process-based models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **55**(1):41–62.

Dunstan P, Foster S, Hui F, Warton D. 2013. Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics* **18**(3):357–375.

Dunstan PK, Foster SD, Darnell R. 2011. Model based grouping of species across environmental gradient. *Ecological Modelling* **222**(4):955–963.

Favrichon V. 1994. Classification des espèces arborées en groupes fonctionnels en vue de la réalisation d'un modèle de dynamique de peuplement en forêt guyanaise classification of guiana forest tree species into functional groups for a model of vegetation dynamic. *Revue d'écologie* **49**:379–403.

Favrichon V. 1998. Modeling the dynamics and species composition of tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes. *Forest Science* **44**(1):113–124.

Fieberg J, Ellner S. 2001. Stochastic matrix models for conservation and management: a comparative review of methods. *Ecology Letters* **4**(3):244–266.

Flores O, Rossi V, Mortier M. 2009. Autocorrelation offsets zero-inflation in models of tropical saplings density. *Ecological Modeling* **220**(15):1797–1809.

Friedman J. 2000. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**(5):1189–1232.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1):1–22.

García-López J, Allué C. 2011. Modelling phytoclimatic versatility as a large scale indicator of adaptive capacity to climate change in forest ecosystems. *Ecological Modelling* **222**(8):1436–1447.

Gitay H, Noble I. 1997. *What are Functional Types and How Should We Seek Them?* Cambridge University Press: Cambridge,3–19.

Gourlet-Fleury S, Blanc L, Picard N, Sist P, Dick J, Nasi R, Swaine MD, Forni E. 2005. Grouping species for predicting mixed tropical forest dynamics: looking for a strategy. *Annals of Forest Science* **62**(8):785–796.

Grün B, Leisch F. 2007. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis* **51**:5247–5252.

Grün B, Leisch F. 2008. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* **28**(4):1–35.

Gupta M, Ibrahim J. 2007. Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association* **102**(479):867–880.
Hargrove W, Hoffman F. 2004. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management* **34**(1):39–60.
He H, Garcia E. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9):1263–1284.
Hérault B, Bachelot B, Poorter L, Rossi V, Bongers F, Chave J, Paine C, Wagner F, Baraloto C. 2011. Functional traits shape ontogenetic growth trajectories among rain forest tree species. *Journal of Ecology* **99**(6):1431–1440.
Hui FC, Warton DI, Foster SD, Dunstan PK. 2013. To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. *Ecology* **94**(9):1913–1919.
Jensen AL. 1996. Density-dependent matrix yield equation for optimal harvest of age-structured wildlife populations. *Ecological Modelling* **88**(1–3): 125–132.
Khalili A, Chen J. 2007. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102**(479):1025–1038.
Leisch F. 2004. FlexMix: a general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* **11**(8):1–18.
Liang J. 2010. Dynamics and management of Alaska boreal forest: an all-aged multi-species matrix stand growth model. *Forest Ecology and Management* **260**(4):491–501.
Liang J, Zhou M, Verbyla D, Zhang L, Springsteen AL, Malone T. 2011. Mapping forest dynamics under climate change: a matrix model. *Forest Ecology and Management* **262**(12):2250–2262.
McLachlan G, Krishnan T. 2008. *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics. Wiley Series in Probability and Statistics: New York.
Monni S, Tadesse M. 2009. A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis* **4**(3):413–436.
Morris WF, Doak DF. 2002. *Quantitative Conservation Biology: Theory and Practice of Population Viability Analysis*. Sinauer Associates, Inc.: Sunderland, MA, 480 pp.
Mortier F, Rossi V, Guillot G, Gourlet-Fleury S, Picard N. 2013. Population dynamics of species-rich ecosystems: the mixture of matrix population models approach. *Methods in Ecology and Evolution* **4**(4):316–326.
Ouédraogo DY, Mortier F, Gourlet-Fleury S, Freycon V, Picard N. 2013. Slow-growing species cope best with drought: evidence from long-term measurements in a tropical semi-deciduous moist forest of Central Africa. *Journal of Ecology* **101**(6):1459–1470.
Pastor J, Post W. 1988. Response of northern forests to CO2-induced climate change. *Nature* **334**:55–58.
Pearson R, Dawson T, Berry P, Harrison P. 2002. Species: a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling* **154**(3):289–300.
Picard N, Köhler P, Mortier F, Gourlet-Fleury S. 2012. A comparison of five classifications of species into functional groups in tropical forests of French Guiana. *Ecological Complexity* **11**:75–83.
Picard N, Mortier F, Chagneau P. 2008. Influence of estimators of the vital rates in the stock recovery rate when using matrix models for tropical rainforests. *Ecological Modelling* **214**(2–4):349–360.
Picard N, Ouédraogo DY, Bar-Hen A. 2010. Choosing classes for size projection matrix models. *Ecological Modelling* **221**(19):2270–2279.
Prentice IC, Sykes M, Cramer W. 1993. A simulation model for the transient effects of climate change on forest landscapes. *Ecological Modelling* **65**(1–2):51–70.
R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: Austria.
Rogers-Bennett L, Rogers D. 2006. A semi-empirical growth estimation method for matrix models of endangered species. *Ecological Modelling* **195**(3–4):237–246.
Scheiter S, Higgins SI. 2009. Impacts of climate change on the vegetation of Africa: an adaptive dynamic vegetation modelling approach. *Global Change Biology* **15**(9):2224–2246.
Schelldorfer J, Meier L, Bühlmann P, Winterthur AXA, Zürich ETH. 2014. Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ℓ1-penalization. *Journal of Computational and Graphical Statistics* **23**(2):460–477.
Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**(2):461–464.
Shao G. 1996. Potential impacts of climate change on a mixed broadleaved-Korean pine forest stand: a gap model approach. *Climatic Change* **34**(2):263–268.
Shimatani I, Kubota Y, Araki K, Aikawa SI, Manabe T. 2007. Matrix models using fine size classes and their application to the population dynamics of tree species: Bayesian non-parametric estimation. *Plant Species Biology* **22**(3):175–190.
Shugart H, West D. 1980. Forest succession models. *BioScience* **30**(5):308–313.
Solomon A. 1986. Transient response of forests to CO2-induced climate change: simulation modeling experiments in eastern North America. *Oecologia* **68**(4):567–579.
Städler N, Bühlmann P, Van De Geer S. 2010. ℓ1-penalization for mixture regression models. *Test* **19**(2):209–256.
Stankowski P, Parker WH. 2010. Species distribution modelling: does one size fit all? A phytogeographic analysis of Salix in Ontario. *Ecological Modelling* **221**(13–14):1655–1664.
Steneck R, Dethier M. 1994. A functional group approach to the structure of algal-dominated communities. *Oikos* **69**(3):476–498.
Stott I, Townley S, Carslake D, Hodgson D. 2010. On reducibility and ergodicity of population projection matrix models. *Methods in Ecology and Evolution* **1**(3):242–252.
Swaine M, Whitmore T. 1988. On the definition of ecological species groups in tropical rain forests. *Vegetation* **75**(1–2):81–86.
Talkkari A, Kellomäki S, Peltola H. 1999. Bridging a gap between a gap model and a physiological model for calculating the effect of temperature on forest growth under boreal conditions. *Forest Ecology and Management* **119**(1–3):137–150.
Tseng G, Wong W. 2005. Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**(1):10–16.
Turner IM. 2001. *The Ecology of Trees in the Tropical Rain Forest*. Cambridge University Press: Cambridge.
Usher M. 1966. A matrix approach to the management of renewable resources, with special reference to selection forests. *Journal of Applied Ecology* **3**(2):355–367.
Usher M. 1969. A matrix model for forest management. *Journal of Biometric Society* **25**(2):309–315.
Weiskittel A, Hann D, Kershaw J,Jr, Vanclay J. 2011. *Forest Growth and Yield Modeling*. Wiley: Chicester.
Zeide B. 1993. Analysis of growth equations. *Forest Science* **39**(3):594–616.
Zou H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476):1418–1429.
Zuidema P, Jongejans E, Chien P, During H, Schieving F. 2010. Integral projection models for trees: a new parameterization method and a validation of model output. *Journal of Ecology* **98**(2):345–355.

SUPPORTING INFORMATION

Additional information and supplementary material for this article, including R code, are available online at the journal's website.

Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm

Journal of Multivariate Analysis 2013

Journal of Multivariate Analysis 119 (2013) 47–60

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm

X. Bry [a], C. Trottier [a,b,*], T. Verron [c], F. Mortier [d]

[a] Institut de Mathématiques et de Modélisation de Montpellier, Université Montpellier II, Place Eugène Bataillon CC 051 - 34095, Montpellier, France

[b] Université Montpellier III, Route de Mende - 34095, Montpellier, France

[c] ALTADIS, Centre de recherche SCR, 4 rue André Dessaux - 45404, Fleury les Aubrais, France

[d] Cirad, UR B&SEF, Biens et Services des Ecosystèmes Forestiers Tropicaux, Campus International de Baillarguet, TA C-105/D - 34398, Montpellier, France

ARTICLE INFO

Article history:
Received 30 September 2011
Available online 12 April 2013

AMS subject classifications:
62-07
62H25
62J12

Keywords:
Supervised component generalized linear regression
Generalized linear models
PLS regression
Fisher scoring algorithm

ABSTRACT

In the current estimation of a GLM model, the correlation structure of regressors is not used as the basis on which to lean strong predictive dimensions. Looking for linear combinations of regressors that merely maximize the likelihood of the GLM has two major consequences: (1) collinearity of regressors is a factor of estimation instability, and (2) as predictive dimensions may lean on noise, both predictive and explanatory powers of the model are jeopardized. For a single dependent variable, attempts have been made to adapt PLS regression, which solves this problem in the classical Linear Model, to GLM estimation. In this paper, we first discuss the methods thus developed, and then propose a technique, Supervised Component Generalized Linear Regression (SCGLR), that combines PLS regression with GLM estimation in the multivariate context. SCGLR is tested on both simulated and real data.

1. Introduction

1.1. Framework

The framework is that of a multivariate Generalized Linear Model (GLM): a set of q random variables $Y = \{y^1, \ldots, y^q\}$ (referred to as "responses") is assumed to be dependent on p common explanatory variables, $\{x^1, \ldots, x^p\}$. Each y^k is modeled through a GLM taking $X = \{x^1, \ldots, x^p\}$ as regressors. Moreover, $\{y^1, \ldots, y^q\}$ are assumed independent conditional on X. All variables are measured on the same n statistical units. The assumption of conditional independence means that the statistical link between the responses is due to their common explanatory variables *only*. In our application on real data (cf. Section 7), we aim at predicting the presence/absence of $q = 10$ common tree species of the Congo Basin rainforests measured on $n = 3000$ plots in the Central African Republic. Y is thus a set of 10 binary variables. We use $p = 46$ environmental regressors reflecting the climate, topography, location, stand structure and photosynthetic activity of each plot. One key point is that we are interested in explanatory structures common to part or all of the y^k's. Another key point is that we

* Corresponding author at: Institut de Mathématiques et de Modélisation de Montpellier, Université Montpellier II, Place Eugène Bataillon CC 051 - 34095, Montpellier, France.

E-mail addresses: bry@math.univ-montp2.fr (X. Bry), catherine.trottier@univ-montp2.fr, trottier@math.univ-montp2.fr (C. Trottier), thomas.verron@fr.imptob.com (T. Verron), frederic.mortier@cirad.fr (F. Mortier).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.03.013

want to be able to deal with many and possibly correlated regressors, so that efficient dimension reduction is needed in the regressor space. We may think of other typical problems: modeling q Poisson-distributed event counts (e.g. failures by type of failure) in a complex system as a function of structural characteristics of the system; modeling q random survival times per unit (e.g. lags between stages of a disease in epidemiology) as a function of the unit's characteristics, etc.

The standard estimation of a GLM maximizes the model fit on all linear combinations of regressors. Doing so, it attaches the same importance *a priori* to linear combinations close to many observed variables (i.e. dimensions that focused a lot of the attention and measuring effort) than to linear combinations far from any of them (i.e. related to weak dimensions of measurement, not to say noise). Take the extreme case where all regressors are highly correlated because they reflect the same latent variable with independent error terms and suppose this latent variable is rather poorly related to the dependent ones. Combining the regressors, one may generate as many noise dimensions. These dimensions may even span a space large enough to provide a model with an excellent fit, although there is but one poorly explanatory structural dimension in regressors. Another way of looking at the contradiction is as follows. On the one hand, such a situation as previously described is known to cause instability of coefficient estimation. On the other hand, the presence of such correlated regressors indicates a major concern as to measuring a single predictive dimension; so, if this dimension were directly observed, and the model were based on it, there would be a single precisely estimated coefficient. In most practical situations, explanatory dimensions are not identified well enough to be measured each through a single variable. So, several indirect measures have to be included into the regressors for each such dimension. This yields many and highly correlated regressors. It is possible to perform some PCA on regressors in order to capture a few uncorrelated principal components (PC's) accounting for a sufficient part of the regressors' information, and use these components as new regressors for the GLM estimation. This Principal Components Generalized Linear Regression (PCGLR) has one possible drawback: PC's optimally capture the information of X *per se*, but not chiefly the information most useful to predict Y.

In order to direct the calculation of components towards the prediction of Y in the classical linear model, PLS Regression (PLSR) currently maximizes a covariance criterion that combines the model's goodness of fit index (R^2) with the variance of the linear combination of regressors, that measures its structural strength. Doing so, PLSR draws this combination towards strong measurement dimensions, i.e. away from structurally weak ones. In the classical linear model framework, PLSR is a successful alternative to PCR (Principal Component Regression). In PLSR just as in PCR, components are definite linear combinations of the x's. Regressing Y on components yields a prediction formula that can then be expressed in terms of the x's. Both methods are a way of regularizing regression, in that they drastically limit the transfer of effects between the x's. PLSR performs better than PCR because it takes Y into account when calculating components. But the PLSR criterion is naturally adapted to the linear context, and not to the GLM one.

There have been attempts to combine PLSR with a GLM. Let us briefly review three of them.

When there is but one response y to be modeled, Marx [4] has proposed an Iteratively Reweighted Partial Least Squares (IRPLS) estimation for Generalized Linear Regression. The principle is based on the fact that the maximum likelihood estimation of a GLM can be carried out by an iterative reweighted least squares (IRLS) procedure [5], derived from the Fisher Scoring Algorithm (FSA). Each iteration of it performs Generalized Least Squares (GLS) using a weighting matrix, the design of which derives from the model's hypotheses, and, as such, depends on the model parameters. Therefore, this weighting matrix has to be updated on every GLS step using the current estimated value of these parameters. Now, the GLS step can be straightforwardly replaced by a PLSR step using the current weighting matrix. This method is consistent both with the linear aspect of PLSR and with likelihood estimation of the GLM, because the weighting matrix deriving from the GLM's likelihood is taken into account in the local PLSR estimation. But this method has not yet been extended to multiple responses.

Following that line and for want of any better method, it could seem handy to deal with multiple responses $\{y^1, \ldots, y^q\}$ by first performing IRPLS with each y^k separately, getting a specific predictor component g^k, then performing PCA on $\{g^1, \ldots, g^q\}$ and taking their first PC f^1 as the overall first predictor component. However, this f^1 would be more of a structure common to separate predictor components than a common predictor component and, even if they may not be far apart in many cases, there is some difference between the two. On the one hand, there clearly is a difference in the variance structure used for estimation: when determining separately the predictor component g^k of y^k, the variance matrix W_k used iteratively is determined by this component which is unconstrained by the other $y^{k'}$s. By contrast, calculating a common predictor component should use variance matrices determined by this component, which is constrained by *all* y^k's. On the other hand, it can be shown, in the classical context of linear modeling, that PCA on multiple separate univariate PLS regressions (PLS1) does not lead to multivariate PLS regression (PLS2). As a consequence, the question of a genuine GLM extension of PLS2 has to be dealt with.

Still in the single y context, Bastien et al. [1] have proposed a different way to extend PLS1 to GLM: PLS Generalized Linear Regression (PLSGLR). PLSGLR is based on the following property: PLS1 of a quantitative variable z on $X = \{x^1, \ldots, x^p\}$ yields a rank 1 component f^1 collinear to the sum of the predictors given by OLS regression of z on each x^j alone. Rank 2 component is obtained likewise after replacing each x^j with its OLS regression residuals on f^1, and so on. Hence an apparently straightforward GLM extension of PLS1: given response y, f^1 of PLSGLR is defined as the standardized sum of predictors given by Generalized Linear Regression (GLR) of y on each x^j alone. What may seem awkward in this extension is the inconsistency in the weighting of observations. Indeed, GLR of y on x^j alone implicitly uses a weighting matrix W_j specific to $(y,\ x^j)$, which is different from the weighting matrix associated with GLR of y on components.

Thus, the estimated variance structure of observations according to the model based on components is never used by this method.

In the multiple y context, Bry [2] has proposed an extension to GLM of Thematic Component Analysis (TCA). As PLS2 is a particular instance of TCA, this method – Generalized Linear Thematic Component Analysis (GLTCA) – also extends PLS2 to GLM. In the particular case of a single group of explanatory variables X predicting Y, f^1 is obtained as follows: (1) GLR of each y^k is performed on X separately, yielding predictor z^k, using its own weighting matrix. Note that, in case of collinearities in X, the set of X's PC's may and must replace X in these GLR's. (2) PLS2 of the z^k's on X is performed, yielding f^1. Indeed, in the context of linear modeling, this exactly yields the first PLS2 component of X. (3) To obtain the rank 2 component, one performs GLR of each y^k on the OLS regression residuals of the x^j's on f^1, together with f^1, in order to deflate the effect of component f^1 in calculating f^2. And so on.

A first asset of GLTCA over PLSGLR is that it deals with multiple responses. In its step 1, GLTCA calculates predictors of each y^k based on the complete X, using the corresponding variance structure, which is both an asset, because a unique and complete model for y^k is estimated with the corresponding weighting system, and a drawback, because this model is likely to be over-adjusted. Besides, modeling the y's separately leads to the same caveats as mentioned above. Step 2 performs pure regularization with uniform weighting structure to find a common predictor component in X. The uniform weighting in this step does not derive from likelihood maximization, but only reflects a default balance of observations in the regularization process. Now, keeping estimation and regularization separate is a major drawback, since the estimated variance of the common component-based regularized GLM does not intervene in its estimation.

These theoretical flaws of PLSGLR and GLTCA are what Supervised Component Generalized Linear Regression (SCGLR) was designed to remedy. Only Marx's IRPLS integrates regularization into the estimation algorithm, ensuring that on each step, the estimated variance structure of the regularized model is used to estimate it. The purpose of SCGLR is to extend IRPLS to the multiple response case.

1.2. Plan of the paper

In Section 2, we recall the PLS2 mechanism. In Section 3, we recall the FSA. In Section 4, we show how to nest PLSR within the FSA, and show how it takes the GLM variance structure into account. Section 5 introduces tuning parameters that make the algorithm more flexible. In Section 6, we study the performance of our algorithm on simulated data structures. We finally apply SCGLR to real data in Section 7.

2. Multivariate PLS regression (PLS2)

2.1. Notations

- A being any matrix, A' = transpose of A.
- M being a symmetric semi-definite positive $d \times d$ matrix: $\forall a, b \in \mathbb{R}^d : \ \langle a|b\rangle_M = a'Mb$ refers to the Euclidean scalar product of a and b with respect to metric M.
- $\forall a_1, \ldots, a_h \in \mathbb{R}^d : \ \langle a_1, \ldots, a_h\rangle$ refers to the space spanned by these vectors.
- A being any matrix, $\langle A\rangle$ denotes the space spanned by the column-vectors of A.
- X being a $n \times p$ matrix and $\mathbb{R}^n$ being endowed with metric W, Π_X denotes the W-orthogonal projector onto $\langle X\rangle$.

2.2. Rank 1 problem and solution

Let $X = \{x^1, \ldots, x^p\}, Y = \{y^1, \ldots, y^q\}, f = Xu, g = Yv$, with $u'u = v'v = 1$. Let W be the weighting matrix of observations. The classical rank 1 program of PLSR is:

$$P(X, Y) : \max_{u'u=1; v'v=1} \langle Xu|Yv\rangle_W.$$

We show in Appendix A(a) that the f solution of P is the same as that of:

$$P'(X, Y) : \max_{u'u=1} \sum_{k=1}^{q} \langle Xu|y^k\rangle_W^2.$$

2.3. Rank 2 and above

Let $X_0 = X$, and let $f^r = X^{r-1}u_r$ be the rank r component. To calculate f^{r+1}, X^{r-1} is regressed on f^r, with respect to weighting W, leading to residuals:

$$X^r = X^{r-1} - \frac{1}{\|f^r\|_W^2} f^r f^{r\prime} W X^{r-1}.$$

Rank $r+1$ component, f^{r+1}, is found solving $P(X^r, Y)$ or $P'(X^r, Y)$.

2.4. An extended problem

- $P'(X, Y)$ may usefully be extended as follows. Let $s \in [0; \infty)$ be a tuning parameter and let W_k be a weighting matrix associated with y^k. Let W be another weighting matrix, reflecting the importance given *a priori* to each unit (when all units are considered equally important, we have thus $W = \frac{1}{n} I$). Consider the program:

$$P''(X, Y, s): \max_{u'(X'WX)^{-s}u=1} \sum_{k=1}^{q} \langle Xu | y^k \rangle^2_{W_k}.$$

  Here, each y^k is being treated with a specific weighting matrix W_k. Let $\Omega = \sum_{k=1}^{q} W_k y^k y^{k\prime} W_k$. Appendix A(b) shows that the solution of $P''(X, Y, s)$ is the unit eigenvector u_1 of $(X'WX)^s X'\Omega X$ associated with the largest eigenvalue. Parameter s allows us to fine-tune the attraction of Xu_1 towards X's principal components (with respect to weighting matrix W). Indeed:
  - $s = 0$ gives back the original constraint $u'u = 1$.
  - When $s \to \infty$, u_1 is the unit eigenvector of $X'WX$ associated with its largest eigenvalue, so $f^1 = Xu_1$ is precisely X's first PC in the PCA of X weighted by W: with infinite attraction, the y^k's no longer play any role in component extraction.
- Some statistical interpretation remains to be given for the criterion of program P''. For all k, y^k will be taken W_k-centered, which means:

$$\forall k : y^k = \Pi_{e^{\perp_k}} y^k,$$

  where $e \in \mathbb{R}^n$ has all components equal to 1 and $\perp_k$ refers to orthogonality with respect to metric W_k. As a consequence, observations in X may be centered on any $a \in \mathbb{R}^p$ (the proof is given in Appendix A(c)):

$$\forall k : \langle Xu | y^k \rangle^2_{W_k} = \langle (X - ea')u | y^k \rangle^2_{W_k} \quad \forall a \in \mathbb{R}^p.$$

  Then:

$$\begin{aligned} \forall k : \langle Xu | y^k \rangle^2_{W_k} &= \langle (X - e\bar{x}^{k\prime})u | y^k \rangle^2_{W_k} \quad \text{where } \bar{x}^k = \frac{1}{e'W_k e} X'W_k e \\ &= \|(X - e\bar{x}^{k\prime})u\|^2_{W_k} \|y^k\|^2_{W_k} \cos^2_{W_k}\left((X - e\bar{x}^{k\prime})u, y^k\right). \end{aligned}$$

  Thus, we find back the classical interpretation of the covariance criterion used by PLS1, as compounding interpretable terms:
  - $\|(X - e\bar{x}^{k\prime})u\|^2_{W_k}$ is the variance of the component. Under constraint $u'u = 1$, it measures the component's structural strength.
  - $\cos^2_{W_k}\left((X - e\bar{x}^{k\prime})u, y^k\right)$ measures the goodness of fit of the regression model of y^k on X.
- Rank 2 (and higher) components are sought orthogonal with respect to W, in exactly the same way as stated in Section 2.3.

3. Structure and estimation of the generalized linear model (GLM)

3.1. Univariate GLM

3.1.1. Definition

Let y_i and $x_i = (x_i^j)_{j=1,\,p}$ respectively be the vector of dependent and explanatory variables for unit i. Conditional to x_i, y_i is assumed distributed according to a model having an exponential structure [6]. The log-likelihood corresponding to the n-sample is thus:

$$L(\delta; y) = \sum_{i=1}^{n} \left(\frac{y_i \delta_i - b(\delta_i)}{a_i(\phi)} + c(y_i, \phi) \right).$$

Let us recall classical results for this structure:

$$\mu_i = E(y_i) = b'(\delta_i) \Rightarrow \delta_i = b'^{-1}(\mu_i)$$

$$\mathrm{Var}(y_i) = a_i(\phi) b''(\delta_i) = a_i(\phi) v(\mu_i) \quad \text{with } v(\mu_i) = b''(b'^{-1}(\mu_i)).$$

Independence of $(y_i)_{i=1,\,n}$ conditional on $(x_i)_{i=1,\,n}$ implies that they have conditional variance matrix:

$$\mathrm{Var}(y) = \mathrm{diag}\left(a_i(\phi) v(\mu_i)\right)_{i=1,n}.$$

We assume that, underlying each variable y_i, is a predictor η_i that is linear in x_i:

$$\eta_i = \alpha + x_i \beta \quad \text{where } \beta \text{ is a } p\text{-coefficient vector.}$$

The linear predictor and the expectation of response are linked through a *link function* g:

$$\forall i : \eta_i = g(\mu_i).$$

3.1.2. Estimation

Derivation of the log-likelihood of the model with respect to β yields:

$$\nabla_{\beta} L = 0 \Leftrightarrow X' W_{\beta}^{-1} \frac{\partial \eta}{\partial \mu}(y - \mu) = 0 \tag{1}$$

with:

$$W_{\beta} = \operatorname{diag}\left(g'(\mu_i)^2 a_i(\phi) v(\mu_i)\right)_{i=1,n},$$

and:

$$\frac{\partial \eta}{\partial \mu} = \operatorname{diag}\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_{i=1,n} = \operatorname{diag}\left(g'(\mu_i)\right)_{i=1,n}.$$

Equation system (1), not linear in β, is solved using the iterative *Fisher scoring algorithm*. On iteration $t+1$:

$$\begin{aligned}
\beta^{[t+1]} &= \beta^{[t]} - \left(E\left[\frac{\partial^2 L}{\partial \beta \partial \beta'}\right]^{[t]}\right)^{-1} \left(\frac{\partial L}{\partial \beta}\right)^{[t]} \\
&= \beta^{[t]} - \left(X' W_{\beta^{[t]}}^{-1} X\right)^{-1} X' W_{\beta^{[t]}}^{-1} \left(\frac{\partial \eta}{\partial \mu}\right)^{[t]} (y - \mu^{[t]}) \\
&= \left(X' W_{\beta^{[t]}}^{-1} X\right)^{-1} X' W_{\beta^{[t]}}^{-1} z_{\beta^{[t]}}
\end{aligned} \tag{2}$$

where:

$$z_{\beta^{[t]}} = X\beta^{[t]} + \left(\frac{\partial \eta}{\partial \mu}\right)^{[t]} (y - \mu^{[t]})$$

Eq. (2) with given $z_{\beta}^{[t]}$ may be interpreted as GLS estimation in the following linear model, on iteration t:

$$M^{[t]} : z_{\beta^{[t]}} = X\beta + \zeta^{[t]}$$

where: $E(\zeta^{[t]}) = 0$; $V(\zeta^{[t]}) = W_{\beta}^{[t]} = g'^2(\mu_t) V(y_t)$.

We shall refer to $M^{[t]}$ as the (current) *linearized model*.

Note: as the 1st order development of g at point μ yields:

$$g(y) \approx g(\mu) + g'(\mu)(y - \mu) = z,$$

we may perform OLSR of $g(y)$ on X, in order to get an initial value $\beta^{[0]}$. When $g(y)$ is not defined owing to zero-values in data, we propose to take:

$$\forall i = 1, \quad n : z_i^{[0]} = g(\alpha y_i + (1 - \alpha)\bar{y}), \quad \text{with } \alpha = 0.95.$$

3.2. Multivariate GLM with common predictor (MGLMCP)

3.2.1. Definition

We are now considering a multivariate approach to GLM (for an overview, see [3]). Assume that several variables $y^1, \ldots, y^q$ depend on the "same" linear predictor (in fact predictors collinear to the same vector Xu), conditional to which they are independent.

$$\forall k = 1, \quad q : \eta^k = \gamma_k X u = X \gamma_k u = X \beta_k.$$

For obvious identification purposes, we impose $u'u = 1$. Let $H = \{\eta_{ik}\}_{i,k}$ be the predictor matrix.

3.2.2. Estimation

In view of the conditional independence assumption, and independence of units, the log-density is:

$$L(Y|H) = \sum_{i=1}^{n} \sum_{k=1}^{q} L_k(y_i^k | \eta_i^k).$$

As a result, the corresponding linearized model in the FSA is:

$$\forall k = 1, q : z_{k\beta_k} = X\beta_k + \zeta^k, \quad \text{with } \beta_k = \gamma_k u$$

where the $\zeta^{k'}$s are independent and $\forall k : \; E(\zeta^k) = 0$; $V(\zeta^k) = W_{\beta_k}$.

The FSA must be altered, owing to u and $\gamma = (\gamma_k)_{k=1,q}$. Indeed, estimation of model $M^{[t]}$ is carried out iterating the following alternated least squares two-step sequence:

(i) Given γ, vector $\left(z^k_{\gamma_k u}\right)_k \in \mathbb{R}^{nq}$ is regressed on matrix $\gamma \otimes X$, with respect to variance matrix $W_\gamma = \text{diag}\left(W_{\gamma_k u}\right)_k$. The resulting coefficient vector $\hat{u}$ is made unit-norm, yielding new u.
(ii) Given Xu, each $z^k_{\gamma_k u}$ is regressed independently on Xu, with respect to variance matrix $W_{\gamma_k u}$, yielding new γ_k.

The fixed point values of u and γ of these iterations are taken as $u^{[t]}$ and $\gamma^{[t]}$.

4. Supervised component generalized linear regression: principle and basic algorithm

The above-mentioned mechanisms can now be inter-woven to form the basic SCGLR algorithm.

4.1. Rank 1 component f^1

The basic principle of the method we propose is simple: on each step of the FSA in the estimation of the MGLMCP, we replace the GLS regression step with a PLS2 one.

To be precise: at step k of the FSA, the regression of the z^js in the MGLMCP is the solution, as far as u is concerned, of several equivalent programs (for simplicity's sake, let us write z^k for $z^k_{\beta_k}$, and W_k for W_{β_k}):

$$Q1: \underset{\gamma, u:u'u=1}{\text{Min}} \sum_k \|z^k - X\gamma_k u\|^2_{W_k} \Leftrightarrow Q2: \underset{u:u'u=1}{\text{Min}} \sum_k \|z^k - \Pi_{Xu} z^k\|^2_{W_k}$$

where

$$\|z^k - \Pi_{Xu} z^k\|^2_{W_k} = \|z^k\|^2_{W_k} \sin^2_{W_k}(z^k, Xu) = \|z^k\|^2_{W_k}(1 - \cos^2_{W_k}(z^k, Xu)).$$

So:

$$Q2 \Leftrightarrow Q3: \max_{u:u'u=1} \sum_k \|z^k\|^2_{W_k} \cos^2_{W_k}(z^k, Xu).$$

We propose, just as is done in PLS2, to introduce now the component's variance into the criterion to be maximized, by currently replacing $Q3$, in the MGLMCP estimation algorithm, with:

$$R = P''(Z, X, 0): \max_{u:u'u=1} \sum_k \|z^k\|^2_{W_k} \cos^2_{W_k}(z^k, Xu) \|Xu\|^2_{W_k}$$

$$\Leftrightarrow \max_{u:u'u=1} \sum_k \langle z^k | Xu \rangle^2_{W_k}. \tag{3}$$

In view of Section 2.4, the current solution $u^{[t]}$ is the unit eigenvector associated with the largest eigenvalue of matrix:

$$X'\Omega^{[t]}X \quad \text{with } \Omega^{[t]} = \sum_{k=1}^{q} W^{[t]}_k z^{k[t]} z^{k[t]\prime} W^{[t]}_k$$

where the $z^{k\prime}$s have been W_k-centered.

N.B. Program R's criterion not being 0-degree homogeneous in $W^{[t]}_k$, it is important that all $W^{[t]}_k$ be currently normalized to unit-sum.

4.2. Rank ≥ 2 components

4.2.1. Orthogonality of components

We shall ensure zero-correlation of components f^k with respect to a given fixed weighting W. Weighting here is not linked to the variance of the responses, since it does not derive from estimation optimality concerns. If all observations are considered equally important, we must take $W = \frac{1}{n} I_n$.

So, let:

$$f^r = X^{r-1}u^r \quad \text{with } X^0 = X \text{ and } \forall r > 0: X^r = \Pi_{\langle f^r \rangle W - \perp} X^{r-1}. \tag{4}$$

4.2.2. Role of every extra component

Every extra component f^r must complement the existing ones $F^{r-1} = \{f^1, \ldots, f^{r-1}\}$ as much as possible. So, as far as f^r is concerned, F^{r-1} must be viewed as a group of covariates. Now:

$$\cos^2_W(z, \langle F^{r-1}, f^r \rangle) = \cos^2_W(z, \langle F^{r-1} \rangle) + \cos^2_W(z, \langle \Pi_{\langle F^{r-1} \rangle W - \perp} f^r \rangle) \tag{5}$$

$$\text{where } \Pi_{\langle F^{r-1} \rangle W - \perp} f^r = \Pi_{\langle F^{r-1} \rangle W - \perp} X^{r-1} u^r = \tilde{X}^{r-1}_W u^r \tag{6}$$

$$\text{with } \tilde{X}^{r-1}_W = X^{r-1} - F^{r-1}(F^{r-1\prime} W F^{r-1})^{-1} F^{r-1\prime} W X^{r-1}.$$

Let us take a look back at the MGLMCP. Suppose that we already have $r-1$ available components for $\{z_k\}_{k=1,q}$ and we want to look for the best possible rth common component. This component should be the solution of the following program (having the form of Q 3):

$$\max_{f^r \in \langle X^{r-1} \rangle} \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k, \langle F^{r-1}, f^r \rangle).$$

According to (5) and (6), this is equivalent to:

$$\max_{u^r} \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k, \tilde{X}^{r-1}_{W_k} u^r)$$

which we propose, as in (3), to replace with:

$$\max_{u^r : u^{r\prime} u^r = 1} \sum_k \langle z_k | \tilde{X}^{r-1}_{W_k} u^r \rangle^2_{W_k}.$$

So, the solution is the unit-norm eigenvector associated with the largest eigenvalue of matrix:

$$\left[\sum_k \tilde{X}^{r-1\prime}_{W_k} W_k z^k z^{k\prime} W_k \tilde{X}^{r-1}_{W_k} \right].$$

4.3. Basic algorithm

The complete algorithm used to calculate a set of R components according to these principles may be found in Appendix B(a) (algorithm $\mathbf{A}_0$).

4.4. Predictive model

Once the components are calculated, they are used to produce a set of coefficients of the original explanatory variables x^j in a predictive model of Y. Components are first expressed as a function of x^j's (cf. Appendix A(d)): $F = XV$.

Then, estimating the GLM of Y on F along with the constant e yields the predictor matrix H:

$$H = ea + FC = ea + XB \quad \text{with } B = VC. \tag{7}$$

N.B. If X has been standardized prior to SCGLR, Appendix A(e) shows how to get the coefficients of the unstandardized X in the model.

4.5. Model selection

Let $\mathbf{M}_r$ denote the model based on r components. Coefficients B of $\mathbf{M}_r$ can be used to predict $E\left(y^k_i | x_i\right)$ for units not used in their calculations. The quality of prediction is measured through the following cross-validation procedure. The observation sample is first divided into two subsamples: CT (for calibration and testing) and V (for validation). Then, CT is subdivided a given number of times into two subsamples: C (calibration sample) and T (test sample). The model coefficients are estimated using C. The estimated model is then applied to every unit in T, yielding an estimation of each $E\left(y^k_i | x_i\right)$. To each y^k, we associate an appropriate measure of error $\varepsilon_k(C, T, \mathbf{M}_r)$. To each binary y^k, for instance, we associate a ROC curve, and take $\varepsilon_k = 1 - S_k$, S_k being the area under the curve (AUROC). The ε_k's are then averaged over all k's and (C, T) pairs, giving $\bar{\varepsilon}(\mathbf{M}_r)$. Let r^* denote the r giving the smallest $\bar{\varepsilon}(\mathbf{M}_r)$. When dealing with simulated data, r^* thus found must be the true number of components. When dealing with real data, we must apply every $\mathbf{M}_r$ on V, and check that r^* still gives the smallest $\bar{\varepsilon}(\mathbf{M}_r)$.

5. SCGLR: an enhanced algorithm

In the GLR of variable y, the FSA may encounter some difficulty of convergence. This is the case when the coefficient vector β is weakly identifiable (e.g. due to near-collinearities in X, which are bound to occur when regressors are too many). Then the sole likelihood maximization is not sufficient, and taking into account the structural strength of the predictor, as does $\mathbf{A}_0$, may remedy this difficulty. Yet, it will provide all the less help as most components Xu have close variances under constraint $u'u = 1$, and so $\mathbf{A}_0$ too may encounter difficulties. Performing PCGLR does not lead to such difficulties, since (1) PC's are easy to calculate even when they have close (yet unequal) variances, and (2) GLM estimation is carried out after component calculation, thus on a set of uncorrelated variables, which lowers the risk of the FSA not converging. In order to enable tuning SCGLR towards PCGLR, we have added two tuning parameters giving flexibility to the combination of PLS and GLM estimation.

5.1. Tuning the attraction of predictors towards principal components

As shown in Section 2.4, we may fine-tune the attraction of the current component towards X's principal components by using $P''(Z, X, s)$ instead of $P''(Z, X, 0)$ in (3) with varying parameter s.

Whenever, starting with $s = 0$, the algorithm does not seem to converge fast enough, we may increase s by some predefined quantity, typically one unit, and re-run the component calculation. Note that when $s \to \infty$, SCGLR gives back PCGLR. In the sequel of this section, we shall refer to extracting the first eigenvector of $A_s = (X'WX)^s X'\Omega X$ as performing a "tuned" PLS step.

5.2. Tuning the rate of the FSA steps with respect to the PLS steps in the combination

We may choose the number of steps of the FSA to be performed in between each tuned PLS step. Informally: given components F, a certain number of FSA steps of Y on F are performed, possibly until convergence, yielding variables z^k and corresponding W_k. Then, the tuned PLS step of z^k's on X updates the components, and so on.

This enables us to eventually get a converging algorithm. Indeed, pushing s far enough, we get components that weakly vary about PC's. Operating on thus "stabilized" and uncorrelated components, the FSA itself is most likely to converge. Such convergence is of course paid for with less freedom for components to adjust the explanatory model.

5.3. Algorithm

The enhanced algorithm may be found in Appendix B(b) (algorithm $\mathbf{A}_1$).

6. Numerical results on simulated data

6.1. Data generation

The less easy data type to deal with is binary variables, for their values are usually never close to their expectation. So, we chose to use binary responses in our simulations, which were carried out as follows. Consider $n = 1000$ units.

- 175 explanatory variables X are simulated so as to be structured around four uncorrelated factors $\{\phi^1, \phi^2, \phi^3, \phi^4\}$. ϕ^1, ϕ^2 are intended to be the true explanatory unobserved factors of the y's. Factors ϕ^3, ϕ^4 are the basis of structures stranger to the models of the y's.
  - Simulation of $\{\phi^1, \phi^2, \phi^3, \phi^4\}$:
    - Simulate a vector γ of 1000 random numbers uniformly distributed on [0; 1] (abbreviated 1000 r.n. $\sim U_{[0;1]}$), and take $\phi^1 =$ standardized γ (abbreviated std(γ)).
    - For $k = 1$ to 3, simulate a vector δ_k of 1000 r.n. $\sim U_{[0;1]}$, take $\delta_k^* = \left(\delta_k - \frac{1}{m}\sum_{m=1}^{k} \phi^m \phi^{m\prime} \delta_k\right)$ and finally $\phi^{k+1} = \text{std}(\delta_k^*)$.

    Thus, we get four uncorrelated standardized factors.

    Let now a be a parameter tuning noise about factors (roughly, the tangent of the semi-angles of the bundles), and ranging from 1/5 (reduced noise) to 2 (important noise).
  - Simulation of a first bundle of 30 variables, X_1, structured around ϕ^1. For $j = 1$ to $p_1 = 30$:
    - Simulate a vector κ^j of 1000 r.n. $\sim U_{[0;1]}$, and take $\epsilon^j = \text{std}(\kappa^j)$.
    - Let $\lambda^j = \epsilon^j + \alpha_j \phi^2$ where $\alpha_j =$ r.n. $\sim U_{[-1/5;+1/5]}$, and $\gamma^j = \text{std}(\lambda^j)$.

      N.B. This step is necessary to inject a bit of ϕ^2 into the variables. Indeed, if their deviations from ϕ^1 were obtained as vectors of random numbers, they would be almost systematically orthogonal to ϕ^2.
    - Let $\xi^j = \phi^1 + a\gamma^j$, and $x^j = \text{std}(\xi^j)$.
  - Likewise, we simulate a second bundle of 20 variables, X_2, structured around ϕ^2, with noise containing a bit of ϕ^1.
  - Finally, we independently simulate two extra bundles of variables, X_3 and X_4, respectively containing 75 and 50 variables, and structured around ϕ^3 and ϕ^4. Note that they are heavier – i.e. contain more variables – than those corresponding to the true explanatory factors. So, the true explanatory structures are hidden not only in noise, but also amongst stronger structures in X.
  - $X = [X_1, X_2, X_3, X_4]$
- Responses Y are simulated as follows:
  - For $k = 1$ to $q = 10$, simulate $y^k \sim B(1, p_k(\phi^1, \phi^2))$ with:

$$\ln\left[\frac{p_k(\phi^1, \phi^2)}{1 - p_k(\phi^1, \phi^2)}\right] = a_{k1}\phi^1 + a_{k2}\phi^2$$

    where, for $h = 1, 2$: $a_{kh} \sim U_{[-\frac{2}{3};+\frac{2}{3}]}$.

For each value of a, we used the simulation scheme 100 times, each time yielding a pair (X, Y). For each such pair, we randomly divided the sample into 2 subsamples: a calibration one (C) and a test one (T). On each C, we ran the estimation procedure asking for 5 components. Estimation was carried out using algorithm $\mathbf{A}_0$ as follows: starting with $s = 0$, if convergence threshold ($\sin^2(f^{k[m]}, f^{k[m+1]}) < 10^{-2}$) could not be reached in less than 50 iterations (most of the time, less than ten were enough), then increment s by 1 and try again. Convergent estimation giving components denoted $(f^1, \ldots, f^5)$, we calculated all square correlations $\{\rho^2(\phi^k, f^l);\ k = 1, 2;\ l = 1, 2, 3\}$.

Table 1
Square correlations between components and simulated factors.

$a = 0.2$	$\overline{\rho^2}(\phi^k, f^l)$	f1	f2	f3	f4	f5	$R^2_{2,2}$
	ϕ_1	0.719	0.211	0.050	0.018	0.000	0.895
	ϕ_2	0.180	0.679	0.126	0.013	0.000	
$a = 0.5$	$\overline{\rho^2}(\phi^k, f^l)$	f1	f2	f3	f4	f5	$R^2_{2,2}$
	ϕ_1	0.708	0.188	0.097	0.008	0.000	0.888
	ϕ_2	0.189	0.708	0.088	0.011	0.000	
$a = 2$	$\overline{\rho^2}(\phi^k, f^l)$	f1	f2	f3	f4	f5	$R^2_{2,2}$
	ϕ_1	0.630	0.183	0.052	0.009	0.004	0.769
	ϕ_2	0.156	0.568	0.077	0.016	0.006	

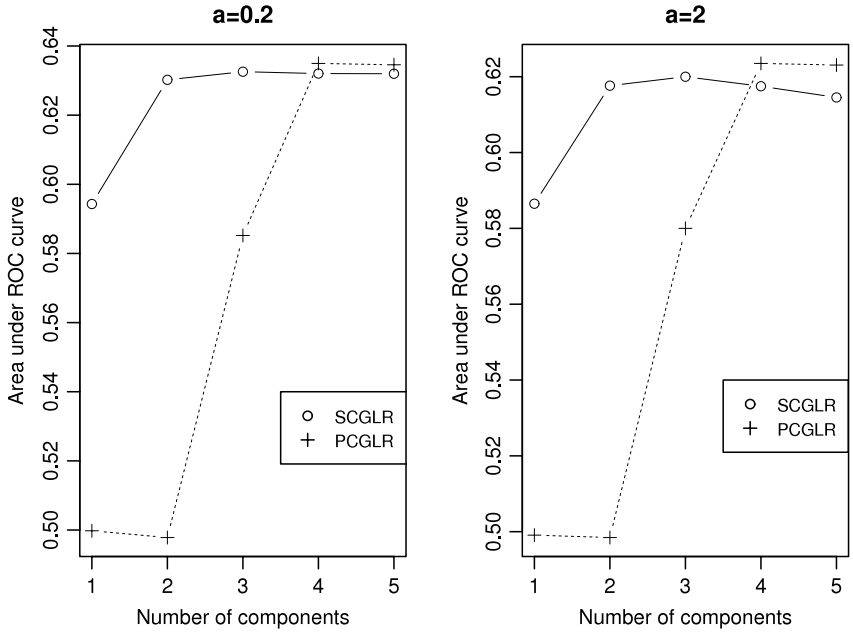


Fig. 1. Area under ROC curves against number of components of simulated data.

For each (X, Y), we considered a sequence of component-based models of the y's, M^k_r denoting the model of y^k based on r components. For each M^k_r, we ran a prediction routine with variable probability threshold $t \in [0; 1]$, yielding a ROC curve, the area under which we calculated (cf. Section 4.5). The same prediction routine was carried out for GLM's based on Principal Components, in order to compare the predictive power of SCGLR with that of PCGLR.

6.2. Results

Convergence was observed in almost all cases with less than 10 steps and $s = 0$ for components f^1 and f^2, which were found highly related to ϕ^1 and ϕ^2. Convergence proved harder for higher rank components, which is obvious, since there are only 2 true predictive factors. When looking for a higher rank component, as SCGLR is no longer led by any goodness of fit, it has to increase s in order to focus on directions closer to the PC's of the residuals of X's regression on the former components.

Components f are more or less drawn towards stronger principal components of X, which have generally no reason to be individually very close to the factors underlying the bundles (unless the latter are uncorrelated, which was the case here). So, these square correlations matter less than their sums: $R^2_{K,L} = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\rho^2(\phi^k, f^l)$. Indeed, $R^2_{K,L}$ close to 1 means that estimation has captured explanatory space $\langle\{\phi^k\}_{k=1,\,K}\rangle$ with component space $\langle\{f^l\}_{l=1,L}\rangle$. What is important is to check that, K being the true number of underlying factors, $R^2_{K,\,K} \approx 1$. This was clearly the case in our simulation, even with the highest degree of noise about factors (cf. Table 1).

So, whether the noise be weak ($a = 0.2$), medium ($a = 0.5$) or strong ($a = 2$), it appears from both Fig. 1 and Table 1 that SCGLR identifies the predictive structures mostly through its first 2 components. PCGLR is fooled at first by the two heavier bundles around ϕ^3 and ϕ^4. Later on, it is able to identify ϕ^1 and ϕ^2 because here these predictive factors are rank 3 and 4 PC's. From the strict standpoint of prediction optimality, Fig. 1 yields $r^* = 3$, but the difference between the areas given by $r = 2$ and $r = 3$ components is so small that component 3 obviously has but a very marginal role.

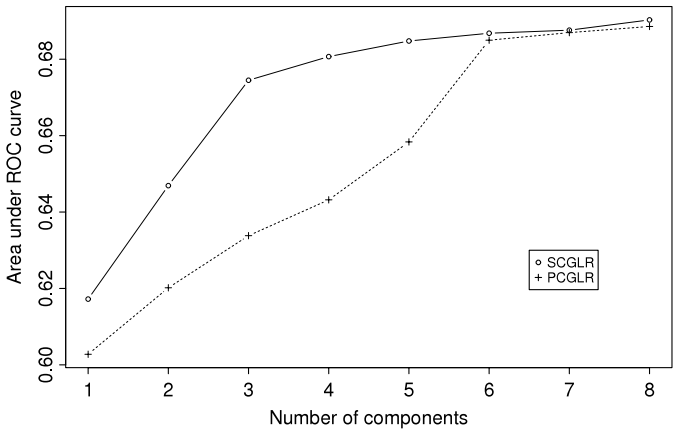


Fig. 2. Area under ROC curves against number of components of simulated data.

Table 2
Percentages of variance of X accounted for by components.

Component rank	1	2	3	4	5	6	7	8
SCGLR (%)	26.6	10.5	10.2	08.1	08.9	03.8	03.8	02.1
SCGLR (% cum.)	26.6	37.2	47.4	55.5	64.4	68.1	71.9	74.0
PCA (%)	28.4	15.1	09.3	07.3	05.5	03.9	03.1	03.0
PCA (% cum.)	28.4	43.5	52.8	60.1	65.6	69.5	72.6	75.6

7. An application to floristic and environmental data

7.1. Data and problem

We aim at predicting the presence/absence of 10 common tree species of the Congo Basin rainforests through environmental variables. The 10 species are mostly timber species, reliably identified in the field. Measures have been obtained on 3000 inventory plots from two logging companies (SCAF and TCA). We assigned to each plot the values of 46 numeric environmental variables reflecting climate, topography, location, stand structure and photosynthetic activity of each plot.

7.2. Results

On the whole sample, SCGLR converged without any difficulty with $s = 0$ for all components except the 7th, for which s was set to 1. The sample has then been randomly divided into 30 subsamples, and each of these has been used in turn as a test sample for prediction, the others being used for calibration. Fig. 2 compares the average AUROC of SCGLR and PCGLR. We can see that SCGLR is at once more efficient for prediction, and is already close to its best performance with $r = 3$ components. PCGLR requires 6 components to reach the same level of performance. SCGLR's graph (respectively PCGLR's) shows a break in the slope at $r = 3$ (respectively $r = 6$), after which the increase becomes very slow, and goes on, so that we can only say $r^* \geq 8$ for both.

Fig. 3 shows the contents of the explanatory space spanned by the first three SCGLR components. Component 1 is illustrated by many variables, and appears to be close to the 1st PC (correlation = 0.94). It essentially sets longitude against correlates of photosynthetic activity. Such an opposition was expected, since eastern forests have a soil which is less rich than that of western ones, and are composed of slower-growing species. Plane (2, 3) produced by SCGLR reveals two explanatory structures. Overall above-ground biomass (agbtot) and total basal area ($g0$) are important quantifiers of competition between trees and environment disturbance. Highlighted by SCGLR's plane (2, 3), they are only captured by the 6th PC. This accounts for the rise in the AUROC of PCGLR on component 6. The bundle orthogonal to agbtot and $g0$ on plane (2, 3) is correlated to latitude (y) and a corresponding North–South climatic gradient. This bundle is rather strongly correlated with the second PC (cf. Fig. 4).

Table 2 provides the percentages of variance of X accounted for by the components of SCGLR and PCA. Indeed, the percentage captured by SCGLR's first 3 components (47.4%) is not so much lower than that of PCA (52.8%), but SCGLR components prove much more predictive.

8. Conclusion

IRPLS being, according to us, the only extension of PLS regression to GLM that respects the variance structure of that model, we have tried to extend it to multivariate responses. In the current GLS step of the Fisher scoring algorithm, we have introduced some multivariate PLS-type regularization. We have bridged our method with PCGLR by introducing a

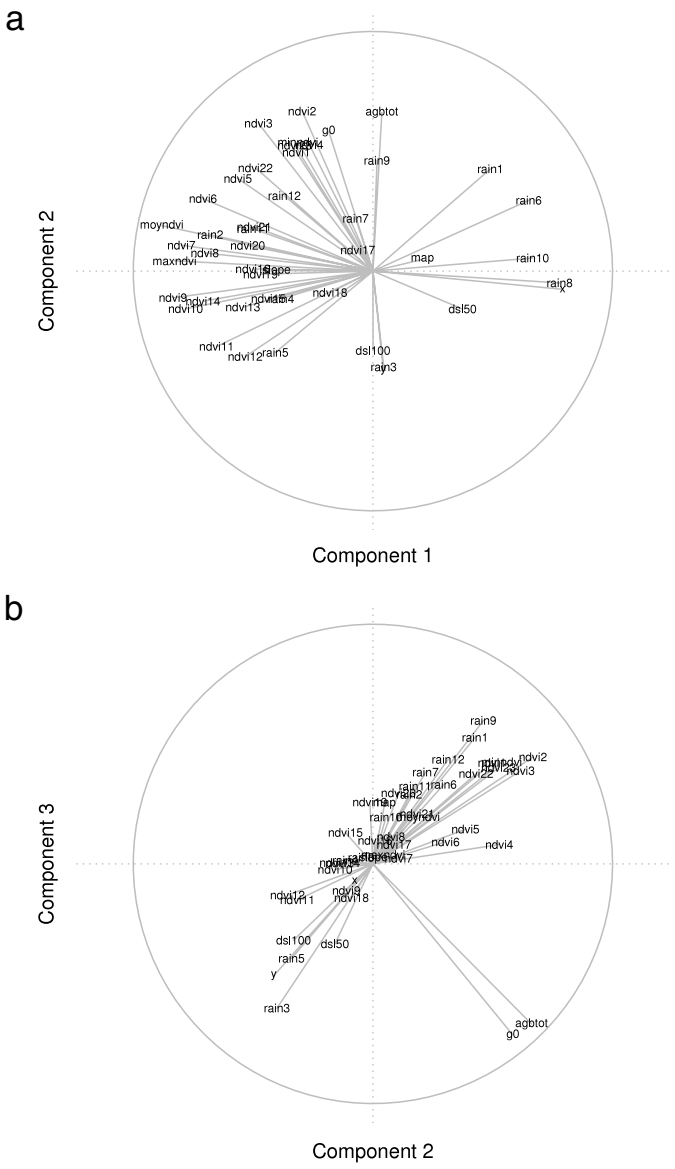


Fig. 3. Correlation scatterplots for SCGLR's first components: (a) components 1 and 2, (b) components 2 and 3.

numeric parameter that allows us to continuously tune the attraction of explanatory components towards the principal components of explanatory variables. The algorithm proved to always converge, and proved able to dig out at once the relevant explanatory and predictive structures, on simulated as well as real data.

Acknowledgments

The forest inventories were funded by the French Agency for Development (AFD) through the PARPAF project. We wish to thank S. Chong (TCA), A. Banos (SCAF), and the "Ministère des Eaux, Forêts, Chasse et Pêche" of the Central African Republic for authorizing access to the inventory data, and the field teams who drew up these inventories. This study is part of the ErA Net BiodivERsA CoForChange project, funded by the National Research Agency (ANR) and the Natural Environment Research Council (NERC), involving 16 European, African and international partners and a number of timber companies (see the list on the website, http://www.coforchange.eu).

Appendix A

(a) Solution of P:

$$L = \langle Xu|Yv\rangle_W - \lambda(u'u - 1) - \mu(v'v - 1)$$

$$\underset{u}{\nabla} L = 0 \Leftrightarrow X'WYv = 2\lambda u \ (1); \qquad \underset{v}{\nabla} L = 0 \Leftrightarrow Y'WXu = 2\mu v \ (1')$$

$$(1, 1') \Rightarrow X'WYY'WXu = \eta u \ (2) \quad \text{and} \quad Y'WXX'WYv = \eta v \ (2') \quad \text{with } \eta = 4\lambda\mu.$$

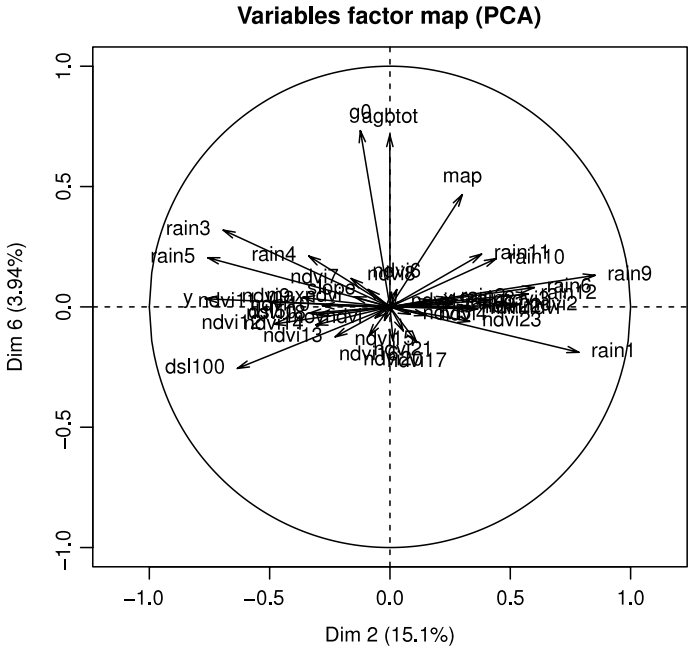


Fig. 4. Correlation scatterplot for principal components 2 and 6.

Besides:

$$u'(1) \Leftrightarrow 2\lambda = u'X'WYv,\ v'(1') \Leftrightarrow 2\mu = u'X'WYv = 2\lambda = \sqrt{\eta} = \langle Xu|Yv\rangle_W$$

which implies that η be maximum. So, solution u is the unit eigenvector u_1 of $X'WYY'WX$ associated with the largest eigenvalue.

Solution of P':

$$\sum_{k=1}^{q}\langle Xu|y^k\rangle_W^2 = \sum_{k=1}^{q}u'X'Wy^ky^{k'}WXu = u'X'W\left(\sum_{k=1}^{q}y^ky^{k'}\right)WXu = u'X'WYY'WXu$$

$$P' : \max_{u'u=1} u'X'WYY'WXu.$$

The solution of P' is given by the unit eigenvector u_1 of $X'WYY'WX$ associated with the largest eigenvalue.

(b)

$$\sum_{k=1}^{q}\langle Xu|y^k\rangle_{W_k}^2 = u'X'\Omega Xu \quad \text{with } \Omega = \sum_{k=1}^{q}W_ky^ky^{k'}W_k$$

$$L = u'X'\Omega Xu - \lambda(u'(X'WX)^{-s}u - 1)$$

$$\nabla_u L = 0 \Leftrightarrow X'\Omega Xu = \lambda(X'WX)^{-s}u \tag{8}$$

$$u'(8) \Rightarrow u'X'\Omega Xu = \lambda, \text{ to be maximized.}$$

$$(8) \Leftrightarrow (X'WX)^sX'\Omega Xu = \lambda u.$$

(c)

$$\begin{aligned}\forall j : \langle Xu|y^k\rangle_{W_k}^2 &= \langle Xu|\Pi_{e^{\perp}k}y^k\rangle_{W_k}^2 = \langle \Pi_{e^{\perp}k}Xu|y^k\rangle_{W_k}^2\\ &= \langle((\Pi_{e^{\perp}k}(X - ea'))u|y^k\rangle_{W_k}^2 = \langle(X - ea')u|\Pi_{e^{\perp}k}y^k\rangle_{W_k}^2\\ &= \langle(X - ea')u|y^k\rangle_{W_k}^2.\end{aligned}$$

(d) We want to write an expression of the form:

$$X^r = X\pi_r. \tag{9}$$

From (4) and (9), we get:

$$f^r = X\pi_{r-1}u^r = Xv^r \quad \text{with } v^r = \pi_{r-1}u^r \tag{10}$$

which leads to:

$$X^r = X\pi_{r-1} - \frac{1}{f^{r\prime}Wf^r} f^r f^{r\prime} WX\pi_{r-1} = X\pi_{r-1} - \frac{1}{f^{r\prime}Wf^r} X\pi_{r-1} u^r f^{r\prime} WX\pi_{r-1}$$
$$= X\left[\mathrm{Id}_p - \frac{1}{f^{r\prime}Wf^r}\pi_{r-1}u^r f^{r\prime}WX\right]\pi_{r-1}.$$

Hence the recurrence formula:

$$\pi_r = \left[\mathrm{Id}_p - \frac{1}{f^{r\prime}Wf^r}\pi_{r-1}u^r f^{r\prime}WX\right]\pi_{r-1}$$

from which we draw $V = [v^1|\ldots|v^R]$ in view of (10).

(e) Let X_o denote the original unstandardized explanatory variable matrix, and X the standardized one. We have:

$$X = (X_o - e(e'We)^{-1}e'WX_o)\Lambda^{-1}, \quad \text{where } \Lambda = \mathrm{diag}(\sigma_k), \qquad \sigma_k^2 = V(x^k)\ \forall k = 1,\ p.$$

So, we have:

$$H = ea + (X_o - e(e'We)^{-1}e'WX_o)\Lambda^{-1}B$$
$$= e(a - (e'We)^{-1}e'WX_o\Lambda^{-1}B) + X_o\Lambda^{-1}B.$$

Hence the model constants: $a - (e'We)^{-1}e'WX_o\Lambda^{-1}B$ and coefficients of variables: $\Lambda^{-1}B$.

Appendix B

(a) **Algorithm $\mathbf{A}_0$**

Initialization

Let: $X^0 = X\ \forall k = 1, q:\ \tilde{X}^0_{W_k} = X$ and $F^0 = \emptyset$

Component iteration

For $r = 1$ to R:

Calculate f^r as follows:

Initialize $Z = [z^1|\ldots|z^q]$ to $Z^{[0]}$ and $\{W_k\}_{k=1,q}$ to $\{W_k^{[0]}\}_{k=1,q} = \{\frac{1}{n}Id_n\}_{k=1,q}$

Iterate from $m = 0$, until convergence:

For $k = 1$ to q:

Standardize every $z^{k[m]}$ with respect to $W_k^{[m]}$

If $r > 1$, set: $\tilde{X}^{r-1}_{W_k^{[m]}} = X^{r-1} - F^{r-1}(F^{r-1\prime}W_k^{[m]}F^{r-1})^{-1}F^{r-1\prime}W_k^{[m]}X^{r-1}$

Define $u_r^{[m]}$ as the unit-norm eigenvector associated with the largest eigenvalue of matrix:

$$\left[\sum_k \tilde{X}^{r-1\prime}_{W_k^{[m]}} W_k^{[m]} z_k^{[m]} z_k^{[m]\prime} W_k^{[m]} \tilde{X}^{r-1}_{W_k^{[m]}}\right]$$

Set $f^{r[m]} = X^{r-1}u_r^{[m]}$

For $k = 1$ to q:

Carry out GLS regression with respect to weighting $W_k^{[m]}$ of each model:

$$z^{k[m]} = \gamma_{k,0} + F^{r-1}[\gamma_{k,1}, \ldots, \gamma_{k,r-1}]' + f^{r[m]}\gamma_{k,r} + \zeta_k$$

thus getting coefficient vector $\gamma_k^{[m]} = (\gamma_{k,0}^{[m]}, \ldots, \gamma_{k,r}^{[m]})$

Update $z^{k[m]}$ and $W_k^{[m]}$ using $\gamma_k^{[m]}$

Set $F^r = [F^{r-1}, f^r]$

Calculate next current X array:

$$X^r = \Pi_{\langle f^r\rangle^{W-\perp}} X^{r-1}$$

(b) **Algorithm $\mathbf{A}_1$**

Initialization

Let: $X^0 = X;\ \forall k = 1, q:\ \tilde{X}^0_{W_k} = X, F^0 = \emptyset$

Component iteration

For $r = 1$ to R:

Calculate f^r as follows:

Initialize $Z = [z^1|\ldots|z^q]$ to $Z^{[0]}$ and $\{W_k\}_{k=1,q}$ to $\{W_k^{[0]}\}_{k=1,\ q} = \{\frac{1}{n}Id_n\}_{k=1,q}$

Iterate from $m = 0$, until convergence:

For $k = 1$ to q:

Standardize every $z^{k[m]}$ with respect to $W_k^{[m]}$

If $r > 1$, set: $\tilde{X}^{r-1}_{W_k^{[m]}} = X^{r-1} - F^{r-1}(F^{r-1\prime}W_k^{[m]}F^{r-1})^{-1}F^{r-1\prime}W_k^{[m]}X^{r-1}$

Define $u_r^{[m]}$ as the unit-norm eigenvector associated with the largest eigenvalue of matrix:

$$(X^{r-1\prime}WX^{r-1})^s \left[\sum_k \tilde{X}^{r-1\prime}_{W_k^{[m]}} W_k^{[m]} z_k^{[m]} z_k^{[m]\prime} W_k^{[m]} \tilde{X}^{r-1}_{W_k^{[m]}}\right]$$

Set: $f^{r[m]} = X^{r-1}u_r^{[m]}$

For $k = 1$ to q:

Set $z^{k[m,1]} = z^{k[m]}$, $W_k^{[m,1]} = W_k^{[m]}$ and $f^{r[m,1]} = f^{r[m]}$

and from $l = 1$ until some convergence precision is reached:

Carry out the current step of the FSA, i.e. GLS regression with respect to weighting $W^{k[m,l]}$ of each model:

$$z^{k[m,l]} = \gamma_{k,0} + F^{r-1}[\gamma_{k,1}, \ldots, \gamma_{k,r-1}]' + f^{r[m,l]}\gamma_{k,r} + \zeta_k$$

thus getting coefficient vector $\gamma_k^{[m,l+1]} = (\gamma_{k,0}^{[m,l+1]}, \ldots, \gamma_{k,r}^{[m,l+1]})$

Update $z^{k[m,l+1]}$ and $W_k^{[m,l+1]}$ using $\gamma_k^{[m,l+1]}$

Update $z^{k[m]} = z^{k[m,\infty]}$, $W_k^{[m]} = W_k^{[m,\infty]}$ and $\gamma_k^{[m]} = \gamma_k^{[m,\infty]}$

Set $F^r = [F^{r-1}, f^r]$

Calculate next current X array:

$$X^r = \Pi_{\langle f^r\rangle W-\perp} X^{r-1}.$$

References

[1] P. Bastien, V. Esposito Vinzi, M. Tenenhaus, PLS generalized linear regression, Computational Statistics and Data Analysis 48 (1) (2005) 17–46.
[2] X. Bry, Extension de l'analyse en composantes thématiques univariée au modèle linéaire généralisé, Revue de Statistique Appliquée 54 (3) (2006).
[3] L. Fahrmeir, G. Tutz, Multivariate Statistical Modeling Based on Generalized Linear Models, Springer-Verlag, New York, USA, 1994.
[4] D. Marx, Iteratively reweighted partial least squares estimation for generalized linear regression, Technometrics 34 (4) (1996) 374–381.
[5] P. McCullagh, J. Nelder, Generalized Linear Models, Chapman and Hall, New York, USA, 1989.
[6] J. Nelder, R. Wedderburn, Generalized linear models, Journal of the Royal Statistical Society: Series A 135 (1972) 370–384.

Résumé Le devenir des forêts est désormais l'une des préoccupations majeures du $20^{\text{ième}}$ siècle. Celles-ci sont justifiées par l'importance que revêtent les forêts pour de multiples acteurs et à de multiples échelles. L'enjeu consiste aujourd'hui à conserver la biodiversité des forêts tropicales et à les gérer durablement, c'est-à-dire à exploiter leurs ressources en préservant à long terme leurs fonctions écologiques, économiques et sociales. Protéger et gérer durablement un écosystème dans son ensemble conduit à le considérer non plus comme un ensemble indépendant de processus biologiques mais comme un ensemble de processus interdépendants. Analyser, comprendre ou encore prédire le future de ces écosystèmes nécessite certaines précautions et des méthodes d'analyses adéquates doivent être employées. C'est ce que je me suis efforcé de faire au cours de ma carrière et ce mémoire, d'habilitation à diriger des recherches, présente les travaux que j'ai été amenés à développer. Il est important de souligner que ce sont les questions biologiques qui ont motivé mes recherches en statistique. Il m'est donc apparu naturel que ce soit au travers des applications que je devais présenter mes activités de recherches en bio-statistiques. La première partie donne un rapide aperçu du contexte biologique et mathématique. La seconde présente plus en détail quatre résultats qui me semblent majeurs et qui traitent de la prise en compte des dépendances spatiales, de la richesse spécifique des écosystèmes tropicaux ou encore des questions de prédictions. La dernière partie présente les stratégies à long terme que je souhaiterais mettre en place pour mener à bien et fédérer les recherches et répondre ainsi à l'objectif commun : la préservation des écosystèmes forestiers compatible avec le développement des populations humaines.